STATISTICAL ISSUES IN THE COMPARISON OF MORTALITY IN NEONATAL INTENSIVE CARE UNITS WITHIN A UK HEALTH REGION

Thesis submitted for the degree of

Doctor of Philosophy

at

The University of Leicester

by

Bradley Neil Manktelow

BSc (Open), MSc (Leicester)

Department of Health Sciences

University of Leicester

November 2005

STATISTICAL ISSUES IN THE COMPARISON OF MORTALITY IN NEONATAL INTENSIVE CARE UNITS WITHIN A UK HEALTH REGION

BRADLEY NEIL MANKTELOW

ABSTRACT

Aims

Since 1990 data have been collected by the Trent Neonatal Survey (TNS) on neonatal intensive care activity within the area of the former Trent Regional Health Authority. While TNS is a unique data set, no systematic investigation had previously been undertaken to ensure that the most appropriate statistical methods were applied to its analysis. In this thesis, methods for the analysis of in-unit mortality rates were reviewed, critically appraised and, where appropriate, developed in order to identify the most suitable methods.

Methods

Statistical methods were illustrated using data from infants born in the years 2000 to 2002, at 32 completed weeks gestational age or less, admitted to one of the sixteen neonatal intensive care units (NICUs) within the area. The methods were discussed and risk-adjustment methods were explored to allow for differences in disease severity between the units.

Results

Simple descriptive approaches and statistical models are presented. In particular, summary statistics derived from logistic regression models were explored, including odds ratios and statistics from both direct and indirect standardization. In the final approach, logistic regression models were applied to obtain estimated standardized mortality ratios (SMRs) for each NICU. Proposed methods to estimate confidence intervals for the SMR were investigated through a simulation study and by application to the TNS data, with the method proposed by Hosmer and Lemeshow (1995) applied in the final models. The use of Bayesian methods was proposed and a model developed allowing the appropriate estimation of all uncertainty.

Conclusions

The use of SMRs was proposed for the reporting of mortality in future TNS annual reports. The advantages of a Bayesian approach, with the ability to make probability statements about the SMR, were also emphasised. Further work is required into the effect of specification of prior distributions before this method can be recommended routinely.

ACKNOWLEDGEMENTS

I thank my supervisors, Professor Keith Abrams and Professor Carol Jagger, for their advice, help and encouragement through these past years.

I am also grateful to the whole TNS team, without whom this thesis would not have been possible. Particular thanks are owed to Elizabeth Draper for all of her help, support and encouragement.

Special thanks must go to Irma for her encouragement and support, and to Lloyd, Xochitl, Lewis and Anna-Maria for their occasional interest.

CONTENTS

TABLES

FIGURES

NOTATION & ABBREVIATIONS

A_{ROC}	Area under the ROC-curve (§6.3.1)
C	Hosmer & Lemeshow goodness-of-fit statistic (§6.3.2)
d_i	$=\begin{cases}0 & \text{if infant } i \text{ discharged alive}\\ 1 & \text{if infant } i \text{ died before discharged}\end{cases}$
$\mathrm{E}(x)$	Expected value of x
$g = \log_e(\omega)$	Log odds of death
$P(x)$	Probability of x
π	Probability of death
$\mathrm{Var}(x)$	Variance of x
z_p	pth percentile of the standard normal distribution
$\psi = \frac{\left(\frac{\pi_1}{1-\pi_1}\right)}{\left(\frac{\pi_2}{1-\pi_2}\right)}$	Ratio of odds of death in group 1 compare to group 2
$\omega = \frac{\pi}{1-\pi}$	Odds of death

BCa	Bias-corrected and accelerated bootstrap (§5.6.3)
BPD	Biparietal diameter
CI	Confidence (or credible) interval
CMF	Comparative Mortality Figure (§5.5.1)
CTG	Cardiotocogramography
EFM	Electronic fetal monitoring
LMP	Last menstrual period
MCMC	Markov Chain Monte Carlo
MLE	Maximum likelihood estimators
NICU	Neonatal Intensive Care Unit
OR	Odds ratio
SMR	Standardized Mortality Ratio (§5.5.2)
TNS	Trent Neonatal Survey (§2.2)

Chapter 1: INTRODUCTION

1.1 Aims of the Thesis

Since 1990, data have been collected, under the auspices of the Trent Neonatal Survey (TNS), on activity in the neonatal intensive care units (NICUs) within the area of the former Trent Health Authority (Derbyshire, Leicestershire, Lincolnshire, Nottinghamshire, South Humberside and South Yorkshire). TNS is a population-based survey of neonatal intensive care provision in the former Trent Health Authority Region, and is described in more detail in §2.2.

In the TNS annual reports risk-adjusted in-unit mortality rates are given for each neonatal unit. The information from TNS forms a unique data set in terms of size, completeness and history, but no systematic investigation has previously been undertaken to ensure that the most appropriate statistical methods are applied to its analysis. This thesis aims to review, critically appraise, and develop where appropriate, possible methods for analysing these data in order to produce the most suitable summary of in-unit mortality, whilst recognising the differing case-mix of the units. The sensitivity of the results to the statistical methodology used is of interest. These methods will be illustrated and developed with data from infants born in the years 2000 to 2002, at 32 completed weeks gestational age or less, and who were admitted to NICUs within the area.

This thesis does not aim to discuss in great depth the rationale for such provider profiling; rather the statistical methodology will be of more interest. However, in order to critically appraise the various statistical methods, some discussion of the wider issues surrounding provider profiling is necessary.

This Chapter introduces the thesis. A brief introduction to provider profiling is given in Section 1.2. Section 1.3 discusses general issues in the process of producing such profiles by considering the three stages: measurement, analysis and action. The subsequent structure of the thesis is described Section 1.4 and Section 1.5 comprises a summary of the Chapter.

1.2 Background to Provider Profiling

There is great interest in comparing the methods and outcomes of health care providers; whether they are individuals such as surgeons or physicians, institutions such as hospitals or wards, or indeed organisations such as primary care trusts. As patients we are keen to obtain the 'best' and most appropriate treatment for us in the hope of obtaining the 'best possible' outcome. As citizens we expect that health care providers make the 'best possible' use of the resources they are given. These are understandable demands that we make on providers and it is equally understandable that we would want to monitor their performance.

Various terms have been used in the medical literature to describe the process of comparing health care providers: for example **ranking** (Spiegelhalter, 2003), **bench marking** (Field *et al,* 2002) and **profiling** (Christiansen and Morris, 1997). Although these terms are often used interchangeably, they imply different emphases in their approaches. The term **ranking** suggests that it is a provider's position (ranking[a]) in some form of league table that is of most importance. As will be shown, it is unlikely that this approach will give much useful information about the performance of a health care provider. **Bench marking**, on the other hand, implies that there is a standard (benchmark[b]) against which institutions can be compared. This may be a useful approach in particular circumstances and will be discussed further in later Chapters. It is felt, however, that the term **profiling**[c] offers a more general description of this activity and, as a consequence, will be the term generally used in this thesis.

In 1998 the United Kingdom Government summarised its rationale for provider profiling in their discussion document on NHS performance (NHS Executive, 1998):

> *"The new approach aims to improve standards of performance across the NHS, and in doing so to tackle the unacceptable variations that currently exist. The way to achieve this is by comparing performance and sharing best practice …".*

---

[a] *"**ranking** ►**noun** a position in a scale of achievement or status"* (The New Oxford English Dictionary, 1998)

[b] *"**benchmark** ►**noun** **1.** a standard or point of reference against which things may be compared or assessed."* (The New Oxford English Dictionary, 1998)

[c] *"**Profiling** ►**noun** [mass noun] the recording and analysing of a person's psychological and behavioural characteristics, so as to assess or predict their capabilities in a certain sphere or to assist in identifying a particular subgroup of people"* (The New Oxford English Dictionary, 1998)

The New York State Department of Health, USA started collecting and publishing information on mortality rates after coronary artery bypass surgery in 1989. By 1998 they had found that the statewide mortality rate had fallen from 3.52 deaths per 100 in 1989 to 2.44 per 100. They credited "*...this significant improvement in patient survival rates in part to the sharing of performance data with hospitals and physicians*" (New York State Department of Health, 1998a). If this were true then it is a powerful argument in support of such profiling and the dissemination of the information.

However, even accepting that this is true, it is not at all obvious how performance can be compared between health service providers. As with any statistical analysis, there is always the danger of providing the 'wrong' answer to your question or, indeed, answering the wrong question altogether. There is also the danger that publication and circulation in the media may mean that the usual caveats and warnings required from statistical analyses may be lost. In the extreme such results *"... may put a veneer of science onto inappropriate statistics."* (Shaw, 1997)

One attempt to describe the characteristics of a useful approach to profiling was also given by Shaw (1997):

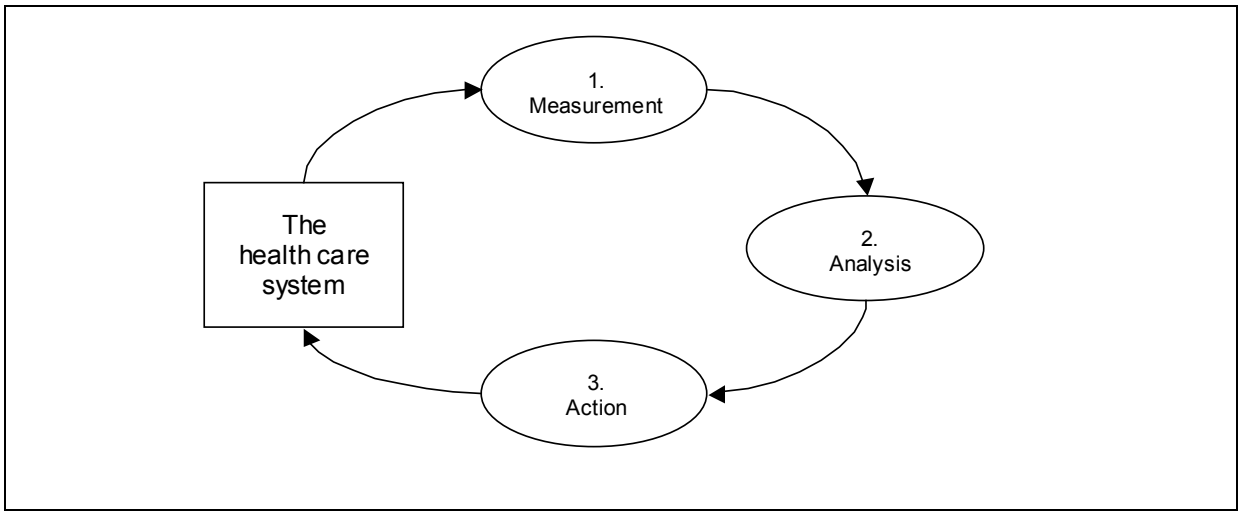
> *"The ideal league table will be based on data which: are specific and sensitive to the achievement of explicit policy targets and clearly measure the achievement of an agreed management objective; represent the legitimate expectations of customers, purchasers and providers (unless these expectations can be reconciled, there will have to be separate indicators for each); are comparable, using consistent definitions and adjustments (they must compare like with like, ensuring comparability of case mix and consistent criteria for numerators and denominators); are accurate, timely and statistically valid (data must be collected promptly, systematically recorded, routinely reported and presented with measures of statistical significance); should assist clinicians and managers to improve performance (they must be readily available for local analysis and comparison to provide incentive for quality improvement rather than perverse incentive for inappropriate activity or manipulation of data); should be cost effective (the staff time and data handling systems required to collect, collate and compare the data should be justified in terms of the quantitative benefits achieved by the indicator); and should enable the public as a whole to assess the service and the individual patient to make informed choices between available technologies and alternative providers."*

How, if at all, such a profile may be achieved is explored further in the rest of this Chapter.

1.3 The Process of Provider Profiling

The process of producing a profile can be divided into three parts: **measurement** (how the performance of the organization is captured in quantitative terms); **analysis** (how the resulting data are interpreted); **action** (the impact of publication on the organization) (Nutley and Smith, 1998). Figure 1.1, adapted from Nutley and Smith (1998), illustrates the model and the process of feedback.

Figure 1.1 A model of the performance measurement process



This thesis will concentrate on the second stage: **analysis**. However, as this cannot be taken in isolation and each of the three stages will be discussed in the rest of this Section.

1.3.1 Measurement

There are many different reasons for the profiling of hospitals, surgeons, doctors and other health care providers. Some of the parties interested in the performance of clinical units are shown in Table 1.1, adapted from Shaw (1979).

However, the question also arises of what constitutes performance and what items in particular are to be measured in order to quantify it. Table 1.2, also adapted from Shaw (1997), offers some suggestions on possible types of measures.

Table 1.1 Performance data – who, why?

Who	**Why**
HM Treasury	Efficiency
Health departments	Public health targets
Politicians	Charter mania
Professions	Self-regulation
Media	Circulation
Purchasers	Contracts
Managers	Accountability
Public	Informed choices

Table 1.2 What is 'performance'?

Access	e.g. waiting times
Process	e.g. immunization, screening
Success	e.g. outcomes
Efficiency	e.g. management costs, output

These needs may often conflict, for example it may not be an efficient use of resources to reduce waiting lists for minor conditions. Therefore, the question of whether a particular set of measures is suitable inevitably, and unsurprisingly, depends on the question being asked, and who is asking it. Anecdotal evidence has been reported (Bagust, 1996) suggesting that the public are interested in questions such as:

> *"Can I get quick and effective local emergency treatment when I need it?*
>
> *If I have a painful and disabling condition and need surgery, how soon can I be treated and get 'back to normal'?*
>
> *In hospital, will I be treated with proper respect and given the individual care and attention I need?*
>
> *Will I be kept informed of everything I need to know about my condition and treatment?"*

These appear to cover different types of measures, with the first, for example, looking at both waiting time and efficacy. To be effective, the choice of indicator should provide a balance between what is measurable and what is important. In addition, practitioners must have the ability to effect change (Jankowski, 1999), although one potential complication is the patient's desire for care and their compliance with treatment (Kassirer, 1994), something practitioners may have limited control over. It has been suggested that the quality of medical care can be assessed according to structure, process or outcome (Donabedian, 1966).

Assessment by structure, process or outcome?

If quality of care is to be measured, it is through the **processes** of medical care that this quality is delivered. Length of stay has been used to compare resource utilization, for example on paediatric intensive care units (Ruttimann and Pollack, 1996), for surgical repair of hip fracture (Shwartz *et al,* 1996) and ICU stay (Knaus *et al,* 1993). However, the measurement of process is difficult, costly and time consuming (Donabedian, 1978). There are difficulties in recording the care received by patients and what aspects of this care were important. This also assumes that there is agreement on the type of care to be given in all circumstances: something which is unlikely to be true except in a very few cases. In neonatal intensive care there is, for example, no consensus on whether infants at the margins of viability should be given aggressive and invasive therapy at all (Greisen, 2004; Levene, 2004). Agreement on the use of particular therapies is likely to be just as difficult to obtain.

The use of **structure** has similar problems. Daley *et al*, for example, argued that the relationship between the presence of a lithotripsy facility, a trauma programme, or a bone marrow transplant programme and the quality of cardiac surgery programme is unproven (Daley *et al,* 1995). There is no evidence that presence of particular neonatal facilities ensures a high level of care, although their absence may mean that poor quality care is provided.

For judging the quality of patient care, measures of **outcome** are often seen as relevant and direct. However, within this are two implicit assumptions. First, that 'good quality' care leads to 'better outcomes' than 'poor quality' care and, second, that rates of adverse outcomes can be used to judge the quality of care given (Thomas *et al,* 1993). The first assumption is likely to hold true through definition: good quality care is that which produces good outcomes (however they are defined). However, it is unclear whether high rates of 'good' outcome necessarily imply that the patients have received good quality care. If they are to prove

satisfactory measures of patient care, they need to adequately reflect the quality of care provided.

Silber *et al* (1995) investigated the use of complication rates as opposed to mortality as a measure of quality of care after coronary artery bypass surgery with over 16,000 patients in 57 US hospitals. They reported that there was low correlation between hospital rankings based on one outcome measure compared to the other. This, they concluded, was evidence that complication rates should not be used to judge hospital care, at least until *"more is known about these differences."* However, there are other conclusions that can be drawn. The assumption made in their work was that mortality is an appropriate measure of the quality of care. It is true that mortality rates are a commonly used measure of performance. On the surface, it certainly appears to be a suitable measure as it is both easily available and relatively reliable, unlike complication rates where different definitions and reporting procedures may exist in different hospitals. However, the ability of death rates to reflect quality of care is not accepted by all (Daley *et al,* 1995; Mant and Hicks, 1995). In fact, Florence Nightingale suggested this in 1863:

> "*The most important perhaps of all the elements are the complications occurring after operations*." (Nightingale, 1863)

The appropriateness of mortality as a surrogate for quality of care may well vary by the medical condition being investigated. A study comparing risk-adjusted mortality rates and peer-review determined quality of care for three conditions admitted to hospitals in Minneapolis and St. Paul, Minnesota found strong support for an association for cardiac disease, equivocal evidence for acute myocardial infarction, and no evidence for septicaemia (Thomas *et al,* 1993).

Even when using death as a measure of poor outcome there are difficulties. For patients with terminal disease, higher death rates are unlikely to reflect poor care and outcomes such as functional status and the quality of the dying experience may be more appropriate measures (Kahn *et al,* 1988). In neonatal care, however, this is unlikely to be the case, but here any decrease in mortality is likely to result in an increased number of infants with severe disabilities (Colver *et al,* 2000; Rijken *et al,* 2003). A more complete picture of outcome can be gained by including long term outcomes (Field *et al,* 2002; Marlow, 2004), but although it is recommended that the health status of survivors be ascertained at at least two years corrected age (that is, two years for the expected date of delivery rather than the actual date of birth) (British Association of Perinatal Medicine, 2001) such data are usually unavailable.

A further difficulty with mortality as the outcome variable to compare hospitals is that the use of hospitals for end-of-life-care varies across areas according to what alternative care is available (e.g. hospice) and this influences the observed hospital mortality rates (Seagroatt and Goldacre, 2004). However, this is not the case in neonatal medicine where care overwhelming takes place on neonatal units.

A quote attributed to Albert Einstein is, *"Everything that can be counted does not necessarily count: everything that counts cannot necessarily be counted."* (World of quotes, 2005). Outcomes such as quality of life and functional status may be more relevant and informative, but collecting such data in a consistent fashion may be extremely difficult (Iezzoni, 1994), especially with neonates. Finally, there is also the assumption made in all profiling that the outcomes monitored are predictable. To be generally accepted any measure used must not only overcome the difficulties discussed above but must also be seen to do so.

However, "*we cannot claim either for the measurement of process or the measurement of outcomes an inherently superior validity compared with the other, since the validity of either follows to an equal degree from the validity of the science that postulates a linkage between the two*." (Donabedian, 1988). Perhaps it is easier to manipulate the figures using process, when compared to outcome. This is directly associated with the former's desirable characteristic that practitioners should be able to effect change. Processes can be changed more easily, but also manipulated more easily. The use of outcomes (or outputs), after allowing for inputs, is the most common approach to profiling and is often referred to as the **input-output** (IO) approach. It is assumed that by comparing the 'output' to the 'input', inferences can be made about the processes that took place in-between even though these have not been directly observed (Draper and Gittoes, 2004).

It seems reasonable that the quality of care provided by a unit cannot be summarised by a single measure. In-unit mortality is only one aspect of neonatal intensive care. A combination of structure, process and outcome can give a more complete picture by simultaneously looking at several markers of neonatal care.

One approach suggested is to use radar charts to map several characteristics together (Leary *et al,* 2002); an example from this paper is shown in Appendix J. In this example, nine axes are shown eminating from a central point, with each axis representing a different parameter. The outcomes for each parameter have been transformed so that a larger value (further from the central point) represents 'better' performance and, thus, a larger area in the polygon represents 'better' overall performance. However, the interpretation of such plots is not straightforward

as the relative importance of each parameter needs to be considered, as does the ordering of the parameters around the plot.

Another possible approach is to consider clinical outcomes together with economic performance (Rapoport *et al,* 1994). The statistical methods discussed and illustrated in this thesis can, with some modification, be used to analyse such a range of measures. An alternative approach may be to use a validated measure of health outcome that can be used to quantify a range of health outcomes into a single value, for example EQ-5D (The EuroQol Group, 1990), SF-36 (SF-36, 2005). However, no such measure currently exists for neonates.

Source and quality of the data

Whatever indicators are chosen, the source of the data used in profiling is, obviously, of great importance. Ideally, such data should to be up-to-date, complete, correct and relevant. Often the data will come from customized datasets (such as TNS), although they may arise from routinely collected data, for example surgeons' logs (Spiegelhalter *et al,* 2002), routine hospital episode data (Dr Foster, 2004), registration of births and deaths (Wen *et al,* 2000). A recent study in France suggested that prospective data collection, that is during the patient's stay, identified more patients with preventable adverse events that the retrospective collection of data from records (Michel *et al,* 2004). Two possible reasons for unreliability are 'confusion' or 'conspiracy' in the hospital administration (Bagust, 1996). It is argued that high priority should be placed on the collection and processing of data by staff trained in research methods. It should also be noted that, in some circumstances, a care provider might feel that it is in its best interest to provide poor figures as a means of putting pressure on a funding or purchasing authority (Bagust, 1996), although direct evidence for this is likely to be hard to find. It may well be that a customised data set would help to reduce the problem of 'confusion'. Evidence from the USA suggests that routinely collected data, usually used by physicians for reimbursement, may be unreliable (Chaiken, 1996). This stems from the very fact that the data are collected for reasons other than provider profiling. The various coding schemes used with such forms (e.g. ICD-9CM and CPT-4) can be difficult to apply and since payment to the physician may not differ greatly between codes there is little incentive for accuracy. It is also noted that where such forms do not need to be submitted for reimbursement, i.e. physicians' payments are capitated, the number of forms submitted is reduced, increasing the likelihood of the data being unreliable.

The introduction of 'payment by results' (Department of Health, 2002b), a fixed price for each patient treated, may supply data that can be used to compare outcomes amongst health

care providers. However, the use of Healthcare Resource Groups (HRGs) to allow for case mix is unlikely to allow adequate adjustment in most cases, and the problems outlined above with using data routinely collected for a different purpose will still hold. The use of routine data, collected for contracting and activity purposes, by the Dr Foster organisation has been criticised for being unreliable (Bridgewater *et al,* 2002).

Poor quality data can produce spurious and inappropriate results (Iezzoni *et al,* 1996b), but the collection of relevant, high-quality data can be hugely expensive. In 1990 the State of California's legislature dropped a bill proposing the collection of clinical information for risk adjustment because of the projected $61.2 million annual cost. The following year they agreed to the use of their existing discharge abstract database with resulting lower costs (Iezzoni, 1997). However, in 1996 a survey reported huge variation in the validity and reporting of risk factors between hospitals in the State of California (Wilson *et al,* 1996). It has been suggested that discharge abstract-based methods are of poorer quality and more open to manipulation through the inclusion of adverse events brought on by substandard care (Iezzoni *et al,* 1996a).

Generating data sets specifically for provider profiling may reduce these potential problems. However, introducing data collection that is additional to routine activities may be difficult in busy medical units. When the Medicare Mortality Predictor Score (MMPS), for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure, was being developed it was estimated that it would take about 15 minutes to abstract the data required from the hospital notes (Daley *et al,* 1988). While computerisation may have aided this task, it is still something that takes a significant amount of time for busy staff. One solution to this problem may lie with electronic data collection (Menke *et al,* 2001). However, there is evidence that automated data collection produces different predictions than the use of manually collected data (Bosman *et al,* 1998). If true, it would, at the least, mean that consistent data collection methods would be required across providers. A national system of (good quality) data collection would allow a more general comparison of NICUs (Field *et al,* 2002; Jain and Fleming, 2004). The Healthcare Commission, following a request from the Department of Health, is currently investigating the possibility of a national ongoing audit of neonatal care in England and Wales (Hubbard and Haines, 2004). However, it is likely that this system, should it be implemented, will only collect a small number of variables. This means that it is unlikely to be sufficiently detailed to allow comparisons of mortality, let alone long-term follow-up.

Reliability is important, although hopefully not too many studies have such a lax approach to data collection as this example (The Guardian, 2003):

> *"British educational leaders with a methodological axe to grind against league tables - isn't that nearly everybody? - will take cold comfort from a university ranking survey published this month by one national Canadian newspaper. The table, carried by the Globe and Mail, awarded high marks to the medical schools at York University and the University of Waterloo, with the latter institution scoring a top 10 place for its law school as well.*
>
> *Unfortunately for the Globe's university report card, as it is known, neither York nor Waterloo has a faculty of medicine, and Waterloo does not offer a law degree.*
>
> *'There is an issue with the overall reliability of the survey,' Nancy White, a spokeswoman for York University, told the Globe and Mail's major national competitor, the National Post."*

1.3.2 Analysis

The second stage of the performance measurement process, outlined in Figure 1.1, is 'analysis'.

Three sources of statistical uncertainty

The statistical uncertainty surrounding reported performance comprises three parts: sampling variability, differences in illness severity between populations and differences in care provision. We wish to quantify the contribution of each source of uncertainty. Indeed, the aim is to account for the first two sources of variability leaving just that due to any difference in the quality of care.

Although the Trent Neonatal Survey aims to record all neonatal intensive care given to infants born at 32 weeks gestational age or less, it is useful to think of these infants as a sample from a (hypothetical) population of infants who might have received intensive care. In this way, the observed mortality rates are assumed to be observed estimates of true (but unknown) underlying mortality rates. This is especially important with the small units.

One definition of statistics is given by Steel and Torrie (1960):

> *"Statistics is the science, pure and applied, of creating, developing, and applying techniques such that the uncertainty of inductive inferences may be evaluated."*

Statistical techniques, used appropriately, quantify this sampling variability and possible methods are outlined in Chapter 3.

The need to adequately adjust outcomes for differences in case mix (**risk-adjustment**) is well recognized (Signorini and Weir, 1999). A hospital or clinician tending to treat only those patients with good prognoses would be expected to have a high rate of 'good' outcome. Conversely those treating patients with poor prognoses would expect a higher rate of 'poor' outcome. Put another way, risk adjustment tries to help answer the question, *"Is it you, Doc, or your patients, who are below average?"* (Poloniecki, 1998). It is important that *"Performance indicators should be measures of what the relevant decision makers can reasonably be held to account for"* (Giuffrida *et al*, 1999). If a unit admits a high proportion of neonates with poor prognoses then this should be taken into account. A study of GP practices has shown evidence that just under half of the variation in hospital admission rates between practices can be accounted for by adjusting for patient characteristics, both socio-demographic and clinical (Reid *et al*, 1999). However, how such adjustment is best achieved is less clear. The methodology for risk adjustment is discussed in greater depth in Chapter 4.

The assumption made is that once sampling variation and case-mix differences have been accounted for, any differences between units is due to differences in the type, or quality, of care that the infants have received. However, what is really 'left' is a measure of what has not been accounted for (Crouchley and Taylor, 2004). This may result from factors other than the quality of care, for example inadequate case-mix adjustment. However, it is hoped that the influence of these other factors is minimal.

Reporting of the results

The results of profiling need to be reported in a usable and useful way to interested parties. Possible approaches for the neonatal mortality data investigated here are taken up in Chapter 5. However, there are three main ways in which the results of provider profiling can be presented:

- The ranks of the institutions (i.e. league tables);
- The probability that the observed outcome is more extreme than could be expected by chance alone, assuming that the null hypothesis is true, for example the institution's true outcome is no different to that of a reference population (i.e. hypothesis testing);

- The probability of the true performance of an institution exceeding, or failing to meet, an agreed clinical standard (i.e. Bayesian posterior probability).

League tables are perhaps the simplest method of presenting the data: care providers are put in order according to the value of their outcome. Information is sometimes presented in this way in the media when reporting the performance of medical or educational institutions: for example, tables in the Sunday Times' Good Hospital Guide (The Sunday Times, 2004b). An example from this guide is shown in Table 1.3.

Table 1.3 Mortality tables from the Sunday Times

HOW THE NHS TRUSTS COMPARE

Ranked by mortality rate from low to high
Average index for Trent Region: 98

! Low ! Average ! High

			Mortality index	Patient satisfaction	Long outpatient waits	Page
!	1	Sherwood Forest Hospitals	85	78%	28%	63
!	2	Sheffield Teaching Hospitals	85	83%	21%	63
!	3	Barnsley District General Hospital	92	74%	27%	61
!	4	Doncaster and Bassetlaw Hospitals	93	80%	10%	61
!	5	Queen's Medical Centre Nottingham University Hospital	97	75%	15%	62
!	6	Nottingham City Hospital	99	82%	12%	62
!	7	University Hospitals of Leicester	100	80%	23%	64
!	8	Rotherham General Hospitals	102	74%	18%	62
!	9	Southern Derbyshire Acute Hospitals	103	78%	17%	64
!	10	United Lincolnshire Hospitals	105	81%	21%	64
!	11	Northern Lincolnshire and Goole Hospitals	108	75%	16%	61
!	12	Chesterfield and North Derbyshire Royal Hospital	110	77%	0%	61

Ranking based on unrounded mortality figures

However, there are serious problems with such an approach. Naturally, there will always be somewhere or someone at the top, and somewhere or someone at the bottom. The ranks themselves give no indication of the size or significance (statistical or clinical) of the differences between the institutions. It is important to distinguish whether a provider is "*indeed an outlier, and not merely 'bottom of the league'* " (Aylin *et al*, 2001a). It has been shown that small changes in outcome can produce substantial changes in rank (Marshall and

Spiegelhalter, 1998b). To quantify this uncertainty in the rankings, it has been suggested that the calculation of confidence intervals (or Bayesian credible intervals) for each rank can illustrate the uncertainty associated with them (Goldstein and Spiegelhalter, 1996; Marshall and Spiegelhalter, 1998b). However, such an approach on its own can be of only limited help as it still offers only indirect information on the size of the clinical differences between the institutions. Such intervals are illustrated in §3.3.1.

By far the most common approach to provider profiling is that using 'classical' hypothesis testing. This is discussed in more detail in §3.2.1 but, in this case, amounts to calculating the probability of obtaining results at least as extreme as those found, if there really was no difference between the units. A problem with such an approach is that a proportion of institutions will always be classified as extreme. Even when all of the health care providers are of similar performance, with such an approach some would still be seen as outliers (Normand *et al*, 1997). It would also be wrong to concentrate solely on statistical significance. While monitoring mortality following coronary artery bypass surgery, the State of New York Department of Health noted that two of the hospitals had mortality rates above the statewide average but that this difference was not statistically significant. However, they concluded that they should still closely monitor their performance (New York State Department of Health, 1998b).

The third approach suggested is to use Bayesian posterior probabilities. The most important difference between this method and the other two is that clinical standards are applied to investigate the performance of the institutions, allowing questions such as *"What is the probability that a given hospital's true mortality rate for cardiac surgery patients exceeded 3.33% last year?"* to be answered (Christiansen and Morris, 1997). This is possible within a Bayesian framework as the parameter is assumed to have a probability distribution as opposed to being of a fixed, but unknown, value (see §3.2.2). The choice of the clinical reference standard will depend on circumstances (Normand *et al,* 1997). It may be an absolute value, such as a national guideline, or a relative measure comparing an institution to a regional average. However, whether the standard is absolute or relative, the reporting of the posterior distribution itself enables readers to decide on the final value of interest.

All of these methods will be illustrated in this thesis. However, when using any of these approaches the final presentation of the data is important. It may be that graphical presentation is the most easy to interpret (Selbmann *et al,* 1982). This was also suggested in an 1857 letter by Florence Nightingale, around the time she produced polar area charts:

> *"which is to affect thro' the eyes what we may fail to convey to the brains of the public through their word proof ears"* (cited in Spiegelhalter, 1999)

She referred to these as 'coxcomb' diagrams, and an example is shown in Appendix J. In these plots, different coloured segments represent the monthly number of deaths from different causes, also allowing temporal trends to be inferred.

However, the most appropriate method of presenting the results is likely to depend to many factors: e.g. the audience, the number of providers, the number of measures. In this thesis tables and simple graphs will be used wherever possible.

1.3.3 Action

The third stage in the performance measurement process is 'action'. The ultimate aim of producing a provider profile is most likely to be to try to influence future events. Such feedback may be structured or haphazard. As Alan Milburn (then Secretary of State for Health) pointed out in 2002, in response to the Kennedy report into paediatric cardiac surgery at Bristol Royal Infirmary (Milburn, 2002):

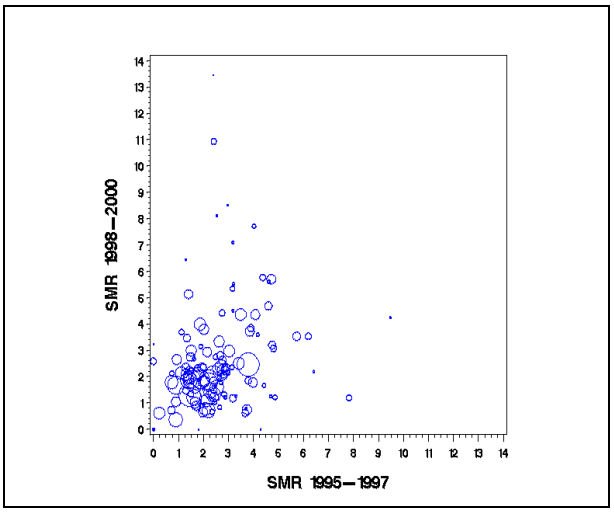
> *"... there was no shortage of data about clinical outcomes at Bristol. The problem was that no one was responsible for analysing them or acting on them."*

Predicting future events

The assumption made is that past events will predict future events; in other words, those providers with a high rate of poor outcomes this year will be those that had a high rate of poor outcomes last year. Green and Wintfeld suggested that the rankings of individual surgeons' risk adjusted mortality rates after coronary-artery bypass grafting in 1989-1991 were poorly correlated with the rankings in 1991-1992: product moment correlation coefficient (ρ) = 0.022 (Green and Wintfeld, 1995). Although there are problems with using the rank to investigate differences (see §3.3.1), such a poor level of correlation may well mean that the assumption of predictive ability is false. However, more recent published data from the New York State Department of Health (New York State Department of Health, 2000; New York State Department of Health, 2004) suggest that such pessimism is misplaced. Using published data, it is possible to compare estimated risk-adjusted mortality rates (more informative than the ranks) following isolated coronary-artery bypass graft, for the years 1995-1997 and 1998-2000, for individual surgeons, practising in the same hospital during both periods, and performing over 10 procedures during each period. In this case ρ = 0.20. This is also

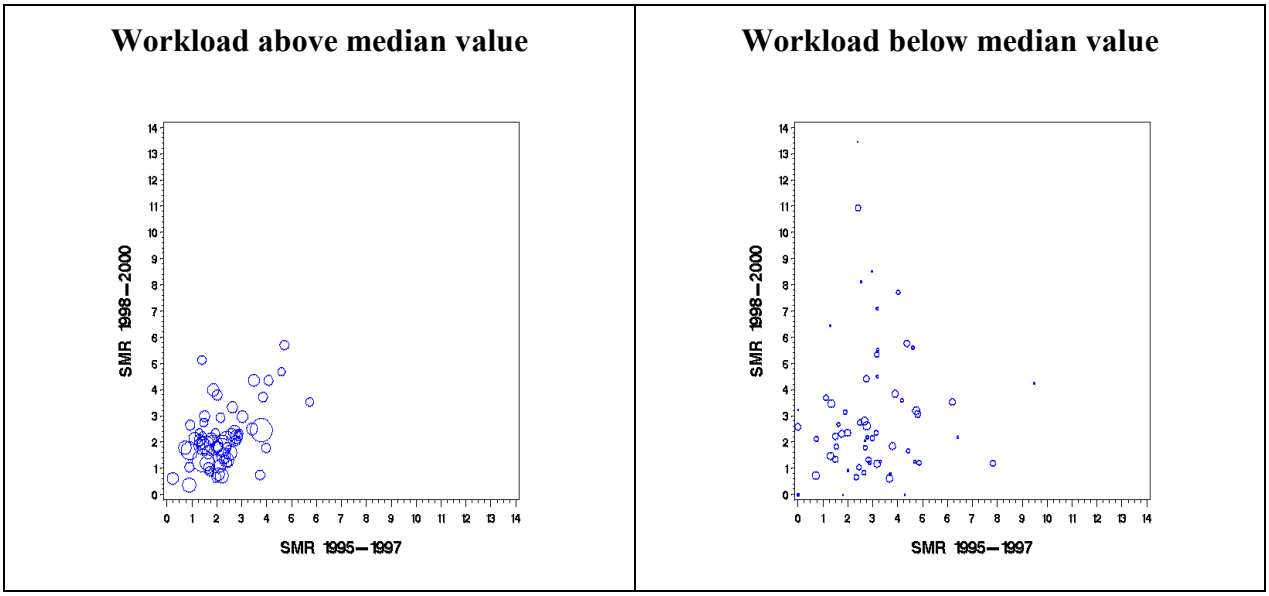
illustrated in Figure 1.2, where the size of the circle is proportional to the average number of procedures carried out by the surgeon.

Figure 1.2 Surgical outcomes in two time periods



In order to investigate whether the differences in the estimated SMRs in the two time periods were the result of random variation, the surgeons are considered in two groups according to whether their workload was above or below the observed median for the whole group (Figure 1.3). For those surgeons with a work-load above the median value, $\rho = 0.48$, whereas for those surgeons with a work-load equal to or below the median value, where the 'true' SMRs are likely to be poorly estimated, $\rho = 0.08$.

Figure 1.3 Surgical outcomes in two time periods by workload



This suggests that where there is sufficient information, there is evidence of correlation between past and future performance.

Reaction of health care providers to profiling

There are further difficulties in securing the confidence of users that profiling can provide useful information. In 1992 reports were first published in Pennsylvania, USA listing annual risk adjusted mortality rates after coronary artery bypass surgery on both hospitals and individual surgeons providing such surgery in the state (Pennsylvania Health Care Cost Containment Council, 1992a; Pennsylvania Health Care Cost Containment Council, 1992b). After the publication of four annual reports, a survey of half of the cardiologists and cardiac surgeons in Pennsylvania reported that 69% (70% cardiologists and 68% cardiac surgeons) of them felt that the reports were "*not important or minimally important*" in assessing the performance of cardiac surgeons (Schneider and Epstein, 1996). There were several concerns reported by respondents to explain this lack of confidence in the reports, the greatest (reported by 79% of responders) being the perceived inadequacy of the risk adjustment methods; although the model used has since been shown to be comparable to other alternative models (Landon *et al,* 1996). A large majority of responders (78%) also felt that the lack of any outcome other that mortality meant that the published figures did not truly reflect the quality of care provided by a surgeon. The third most commonly cited reason for mistrust in the figures was the feeling that surgeons and hospitals could manipulate the figures (53%).

Although it was acknowledged that the responders to this survey (64 cardiologists and 74 cardiac surgeons) may comprise those with stronger, and more negative views, about such public monitoring, this survey shows a large amount of scepticism over the reporting such results. However, despite such scepticism the cardiologists and surgeons felt that the publication of these reports was influencing clinical practice. It was reported that 59% of cardiologists felt that it had become more difficult to find a cardiac surgeon willing to accept a severely ill patient and 63% of the surgeons reported that they were now less willing to operate on such patients. There is a danger that providers may change consciously, or subconsciously, to selecting low-risk patients as a result of previously reported results (Committee on Information Technology in Medicine, 1991). Such avoidance of 'high risk' patients has also been suggested for in-vitro fertilisation clinics (Winston, 1998). It was also suggested in this letter that such clinics may be reluctant to take part in research which they feel may adversely effect their league table position, a concern also expressed in other specialities (Shiu, 2002; Bridgewater *et al,* 2002).

It is possible that even if such data are not made public, biases may be introduced into data collection (Green and Wintfeld, 1995). This is a particular danger as, if it is introduced and

used inappropriately, profiling "*can create an environment of fear instead of fostering quality improvement.*" (Sheldon, 1998). On the other hand, clinicians may welcome the publication of such data if it highlights problems beyond their control, such as resources to enable waiting lists to be reduced (Young, 1993).

How providers react to any review is, unsurprisingly, likely to depend to how their particular organisation fared. When The Times (2001) published the Dr Foster organisation's standardized mortality rates following cardiothoracic surgery, the medical director (Dr Nicholas Bishop) of the trust responsible for the unit claimed to be the best performing, United Bristol Healthcare, said, *"We're delighted to be able to be recognised as being good for heart surgery"* (Vass, 2001). On the other hand, one consultant cardiologist from the hospital with the highest reported death rate, University Hospital of Coventry and Warwickshire, quite rightly points out the lack of adequate risk adjustment (Shiu, 2002) and consultants from the hospital with the second highest rate also point out several deficiencies with the data (Bridgewater *et al,* 2002). It is probably not too cynical to suggest that organisations deemed to be performing 'well' are more likely to turn a blind eye to the deficiencies of any study, while those with 'poor' results will highlight any perceived deficiencies in the analyses, just as, in the early days, work by Florence Nightingale was often poorly received (Iezzoni, 1996). It is important, therefore, to promote such reports as opportunities for improvement, such as has been advocated for reported medical 'near-misses' (Expert Group on Learning from Adverse Events in the NHS, 2000) rather than as metaphorical sticks with which to beat people. It is vital to create the proper climate for good practice to flourish rather than generating fear that poor practice will be discovered and disapproved (Donabedian, 1978). This issue has been bought to the fore more recently by the Institute of Medicine's report *To Err Is Human*. Here, it was argued that it is focusing on systems rather than individuals that allows improvements in the quality of care to be made:

> *"The focus must shift from blaming individuals for past errors to a focus on preventing future errors by designing safety into the system. This does not mean that individuals can be careless. People must still be vigilant and held responsible for their actions. But when an error occurs, blaming an individual does little to make the system safer and prevent someone else from committing the same error."* (Kohn *et al*, 2000: 5)

Such an approach may sit uneasily with the tradition view of medical practitioners, often instilled during their training, that mistakes are not to be tolerated and that the individuals

responsible are to be identified and castigated (Bates and Gawande, 2000; Classen and Kilbridge, 2002). However, *"... simply telling our doctors and nurses to 'try harder' –not to kill their patients by mistake– has nothing at all to do with our eventual success."* (Berwick, 2001). One example of a 'system' change that may potential reduce medical errors is the reduction in working hours for junior doctors, as it is recognised that sleep deprivation can lead to an increase in errors (Clarke, 2001; Feyer, 2001; Pickersgill, 2001).

However, finding causes of error in health care systems, and their solution, is complex requiring a comprehensive approach (Becher and Chassin, 2001; McNutt *et al*, 2002). Provider profiling, as described in this thesis, can play a role in identifying potential areas of concern.

Potential consequences of provider profiling

One example of benefit resulting from the use of profiling results is reported from a study in three New England states: Maine, New Hampshire and Vermont. Using data from coronary artery bypass patients, cardiothoracic surgeons were given feedback on outcomes, training in continuous quality improvement techniques and undertook visits to other hospitals. Over the course of these interventions a 24% reduction in hospital mortality was noted (O'Connor *et al,* 1996). However, no matter how wide the range of measures reported, there is always the danger that efforts on improvement concentrate solely on the reported measures.

The danger with focusing on outliers that perform badly, as may be natural, means that perhaps not enough attention is paid to the units that are producing good results (Sheldon, 1998; Mohammed *et al,* 2001b). This is where good practice is to be found. The largest improvement in the quality of care will occur where the majority of providers with average performance improve: "*if the mean of the quality curve is shifted*" (Scally and Donaldson, 1998). In addition, it is important for health care providers, and those involved in medical care as a whole, to be aware of what is happening:

> *"... even under the best conditions, constant monitoring will have to be maintained, for without it medicine cannot see itself, nor know where it is going."* (Donabedian, 1978)

Ultimately, the question is not whether the results represent the 'truth', as they do not, even assuming that such a truth exists. In the final model, there undoubtedly still remain errors in risk-adjustment, model specification, systematic reporting, coding errors and biases. However, such an analysis is not intended to be a 'final' answer. Instead, its usefulness

should be judged by whether the results are reliable enough, and generates sufficient confidence in practitioners, to inform the debate and choice of actions to improve the quality of neonatal care.

> "*Risk-adjusted mortality rates, therefore, should be supplemented by review of the actual care rendered before conclusions are drawn regarding effectiveness of care.*" (Jencks *et al,* 1988)

The Royal Statistical Society Working Party on Performance Monitoring in the Public Services recently concluded that *"Done badly,* [performance monitoring] *can be very costly and not merely ineffective but harmful and indeed destructive"* and that performance indicators used in monitoring should be seen as *"screening devices"* (Royal Statistical Society Working Party on Performance Monitoring in the Public Services, 2004).

Exactly what is being measured and how it can be interpreted should be carefully defined otherwise confusion can occur, such as the national newspaper headline *(The Times, 2003)*:

> *"Top heart hospital has worst bypass surgery death rate"*

Quite how this qualifies the hospital to be a 'top' hospital is not made clear. The article then proceeds to state *"And its figures for death after aortic valve surgery are the second highest in the country ..."*. This is unlikely to be consistent with any definition of a 'top' hospital.

1.4 Structure of the Thesis

The remainder of this thesis will investigate and discuss statistical issues of particular relevance to the profiling of neonatal intensive care units using in-unit mortality. The data used in this thesis, and their source, the Trent Neonatal Survey (TNS), are introduced and described in Chapter 2, together with relevant background to neonatal intensive care medicine and its organisation.

In Chapter 3, statistical methods that may be appropriate to the profiling of neonatal mortality are described, critically appraised and, where appropriate, illustrated using TNS data. In particular, logistic regression models are introduced.

Chapter 4 sets out the necessity, rationale, and use of risk adjustment and describes published neonatal mortality risk-adjustment scores and their use.

Potentially useful outcome summary statistics derived from logistic regression models are described, illustrated and discussed in Chapter 5. In particular, standardization methods are reviewed, and methods for estimating confidence intervals for the standardized mortality ratio explored through a simulation study and by applying the methods to the TNS data. A Bayesian modelling approach is developed and investigated.

Chapter 6 contains an investigation into potential confounders and effect modifiers that may be associated with in-unit mortality and, therefore, included in any risk-adjustment model. The final model is then developed in Chapter 6, and model fitting and checking techniques are also discussed. Sensitivity analyses are undertaken to explore the robustness of the results.

Chapter 7 contains a discussion of the results from this thesis and the conclusions that can be drawn. Further work following from this thesis is outlined.

The primary statistical software used throughout this thesis will be SAS/STAT®[d] version 8.2, with SAS/BASE® software used for data management. The WinBUGS version 1.4 software (Spiegelhalter *et al,* 1999b) was used for the Gibbs sampling modelling. The STATA® software was used in §6.4.3 for models which included fractional polynomials and the R package (R Development Core Team, 2005) was used in §5.10 to estimate non-linear mixed models with non-Normal mixing distributions. SAS/GRAPH® software was used for all of the Figures, except for Figure 2.1, Figure 2.2 and Figure G.29, which were created with MapInfo Professional® V5.5.

1.5 Chapter Summary

This Chapter introduced the thesis and the process of provider profiling. The aims of the thesis were set out (§1.1), and the background to provider profiling was introduced in §1.2. While §1.3 described the process of profiling, the rest of this thesis will focus on the statistical analyses. The data to be used are introduced in the next Chapter, together with their source, the Trent Neonatal Survey, and a discussion of relevant issues in neonatal intensive care.

[d] SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Chapter 2: THE TRENT NEONATAL SURVEY

2.1 Chapter Overview

This Chapter introduces the data that form the basis of this thesis and their source, the Trent Neonatal Survey (TNS), together with relevant background to neonatal intensive care.

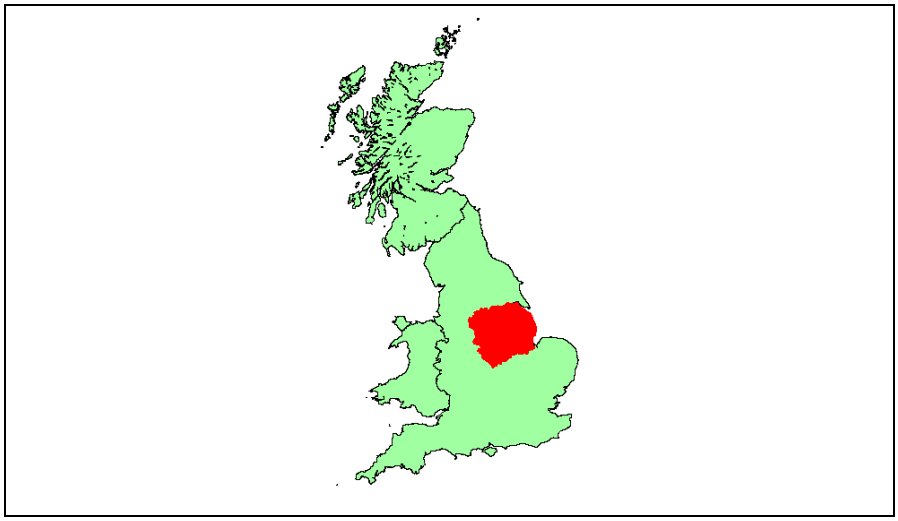
The TNS itself and its development are described in §2.2, whilst §2.3 comprises a brief description of neonatal intensive care, its organisation and issues relevant to neonatal mortality. The TNS data used to illustrate the statistical methods are introduced in §2.4, together with basic summary descriptive information on the infants (more detailed information on these and other variables is given in §6.4 and Appendix G). Methods for ensuring plausible recorded birth weights for gestational age are also introduced and the allocation of transferred infants to a particular unit are discussed. Section 2.5 provides a summary of the main points from the Chapter.

2.2 Trent Neonatal Survey

The Trent Neonatal Survey was first established in 1987 to review the whole neonatal service within the area of the then Trent Regional Health Authority over a one year period (Field *et al,* 1989). The former Trent Region Health Authority comprised the counties of Leicestershire, Rutland, Derbyshire, Nottinghamshire, Lincolnshire, South Yorkshire and South Humberside (Figure 2.1). This region has a population of some 4.6 million with around 60,000 births per annum. In 2002 the neonatal death rate (deaths before four weeks of life) was 4.2 per 1000 live births compared to 3.6 for England and Wales as a whole.

Although the original study lasted only one year, data collection recommenced in February 1990 and has continued since that time. The original exercise in 1987 collected data on every admission to a neonatal unit in Trent. However, the analysis of those data revealed that the majority of admissions were of mature infants for short duration stays and who did not require intensive care.

Figure 2.1 Former Trent Health Authority



(This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown.)

When the survey was re-established in 1990, it was decided to concentrate on those infants identified in the original study as being the most labour intensive. As a result, infants are included in TNS only if they meet at least one of the following inclusion criteria (The Trent Infant Mortality and Morbidity Studies, 2003):

- less than or equal to 32^{+6} weeks gestational age [e];
- less than or equal to 1500 grams birth weight;
- involved in transfers;
- receive any intensive care;
- die in a neonatal unit;
- at term, show signs of severe hypoxic ischaemic encephalopathy.

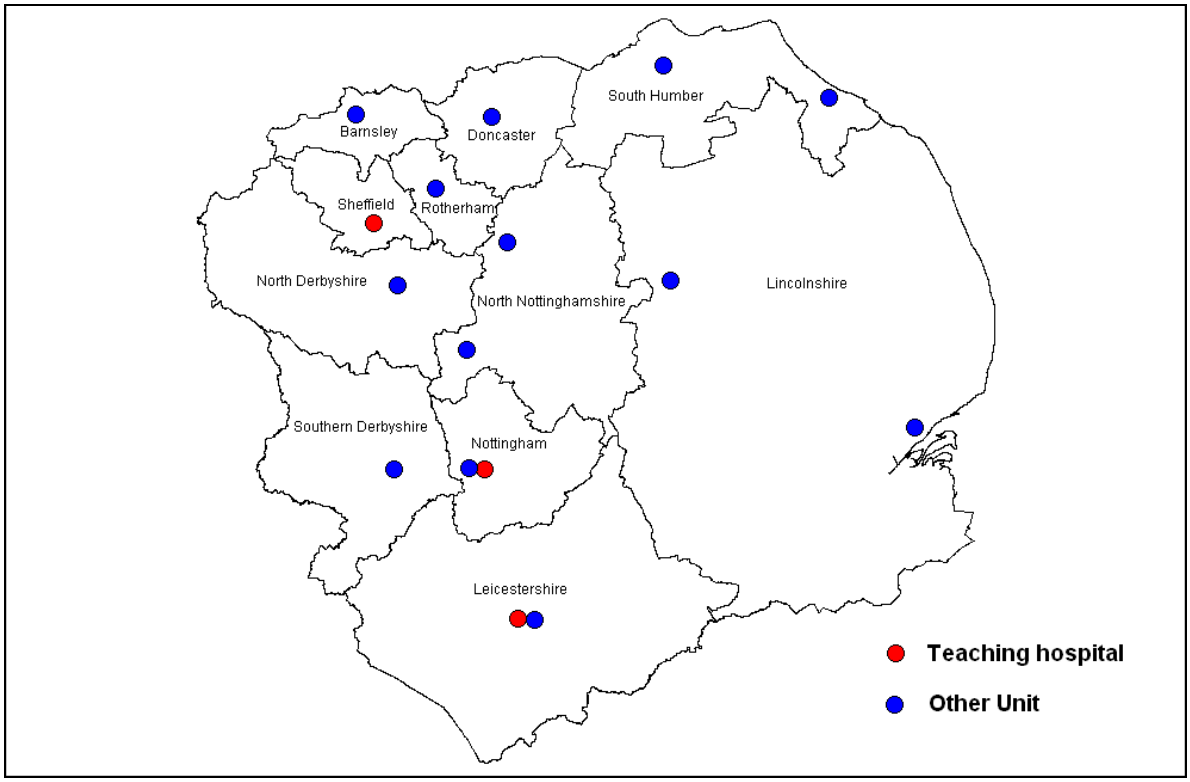
During the time the data used in this thesis were collected there were 16 neonatal intensive care units within the Region (Figure 2.2), located within the following hospitals:

- Barnsley District General Hospital
- Rotherham District General Hospital
- Doncaster Royal Infirmary
- Chesterfield and North Derbyshire Royal Hospital

[e] Gestational age given as $WEEKS^{+DAYS}$

- Jessop Wing, Sheffield (a single unit following the merger of Jessop Hospital for Women and Northern General Hospital in 2001 and considered as a single unit for this thesis)
- Bassetlaw District General Hospital, Worksop
- Kings Mill Hospital, Sutton-in-Ashfield
- Derbyshire Children's Hospital, Derby
- Nottingham City Hospital
- Queen's Medical Centre, Nottingham
- Lincoln County Hospital
- Pilgrim Hospital, Boston
- Leicester Royal Infirmary
- Leicester General Hospital
- Diana, Princess of Wales Hospital, Grimsby
- Scunthorpe General Hospital

Figure 2.2 NICUs in the former Trent Regional Health Authority



Three of these units are in teaching hospitals: Jessop, Queen's Medical Centre and Leicester Royal Infirmary. These units act as referral centres for the other units. In this thesis, the NICUs have been anonymized following current practice in the Trent Neonatal Survey Annual Report (The Trent Infant Mortality and Morbidity Studies, 2003).

Data for the Trent Neonatal Survey are collected by five part-time research nurses who visit each of the neonatal units on a regular basis and complete a standardized data set for each infant. Information is obtained from the clinical records, discussions with staff and, where appropriate, personal observation. A questionnaire is then completed for each admission (Appendix A) and the data double entered to a specifically created Microsoft Access database at the Department of Health Sciences, University of Leicester. Data verification methods comprise range checks and sample validation. The research nurses also meet regularly to discuss difficulties and to ensure the agreed definition of variables.

Funding was originally provided from each of the eleven Health Districts in Trent, with Leicestershire Health acting as the lead agency. Since 2002 funding has been provided by the Primary Care Trusts (PCTs) covered by the survey, with Eastern Leicestershire PCT acting as lead. During the time of data collection for this thesis the Trent Neonatal Survey was overseen by the Trent Institute for Health Services Research, which itself was a collaboration of Leicester & Warwick, Nottingham and Sheffield medical schools. In 2004 the survey was expanded to include data collection from Northamptonshire and the whole of Yorkshire.

The Trent Neonatal Survey is a unique data set due to the length of time it has been running, because the data are collected by specially trained neonatal nurses and also because of its size. The TNS is collected over a complete geographical region, meaning that it is population based and is not subject to the referral biases that arise from only using admissions to individual units. Information from the survey is disseminated through an annual report, an annual research meeting, requests from participating Trusts and Units, conference presentations and the publication of peer-review papers in scientific journals. The survey team participate in national and international collaborative work using TNS data.

2.3 Neonatal Intensive Care

The British Association of Perinatal Medicine (BAPM) define levels of neonatal care as follows (British Association of Perinatal Medicine, 2001:13-14):

"Intensive Care

These babies have the most complex problems. They need 1:1 care by a nurse with a neonatal qualification. The possibility of acute deterioration is such that there should be the constant availability of a competent doctor.

1 *receiving any respiratory support via a tracheal tube and in the first 24 hours after its withdrawal*

2 *receiving NCPAP* [Nasal Continuous Positive Airway Pressure] *for any part of the day and less than five days old*

3 *below 1000g current weight and receiving NCPAP for any part of the day and for 24 hours after withdrawal*

4 *less than 29 weeks gestational age and less than 48 hours old*

5 *requiring major emergency surgery, for the pre-operative period and post-operatively for 24 hours*

6 *requiring complex clinical procedures:*

- *Full exchange transfusion*
- *Peritoneal dialysis*
- *Infusion of an inotrope, pulmonary vasodilator or prostaglandin and for 24 hours afterwards*

7 *any other very unstable baby considered by the nurse-in-charge to need 1:1 nursing: for audit, a register should be kept of the clinical details of babies recorded in this category*

8 *a baby on the day of death.*

High Dependency Care

A nurse should not be responsible for the care of more than two babies in this category -

1 *receiving NCPAP for any part of the day and not fulfilling any of the criteria for intensive care*

2 *below 1000g current weight and not fulfilling any of the criteria for intensive care*

3 *receiving parenteral nutrition*

4 *having convulsions*

5 *receiving oxygen therapy and below 1500g current weight*

6 *requiring treatment for neonatal abstinence syndrome*

7 *requiring specified procedures that do not fulfil any criteria for intensive care:*

- *Care of an intra-arterial catheter or chest drain*
- *Partial exchange transfusion*
- *Tracheostomy care until supervised by a parent*

8 *requiring frequent stimulation for severe apnoea.*

Special Care

A nurse should not be responsible for the care of more than four babies receiving Special or Normal Care.

Special care is provided for all other babies who could not reasonably be expected to be looked after at home by their mother.

Normal Care

Is provided for babies who themselves have no medical indication to be in hospital."

Using these definitions the BAPM has further defined three levels of neonatal intensive care units (NICUs) (British Association of Perinatal Medicine, 2001:2):

"Level 1 *Units provide Special Care but do not aim to provide any continuing High Dependency or Intensive Care. This term includes units with and without resident medical staff.*

Level 2 *Units provide High Dependency Care and some short-term Intensive Care as agreed within the Network.*

Level 3 *Units provide the whole range of medical neonatal care but not necessarily all specialist services such as neonatal surgery."*

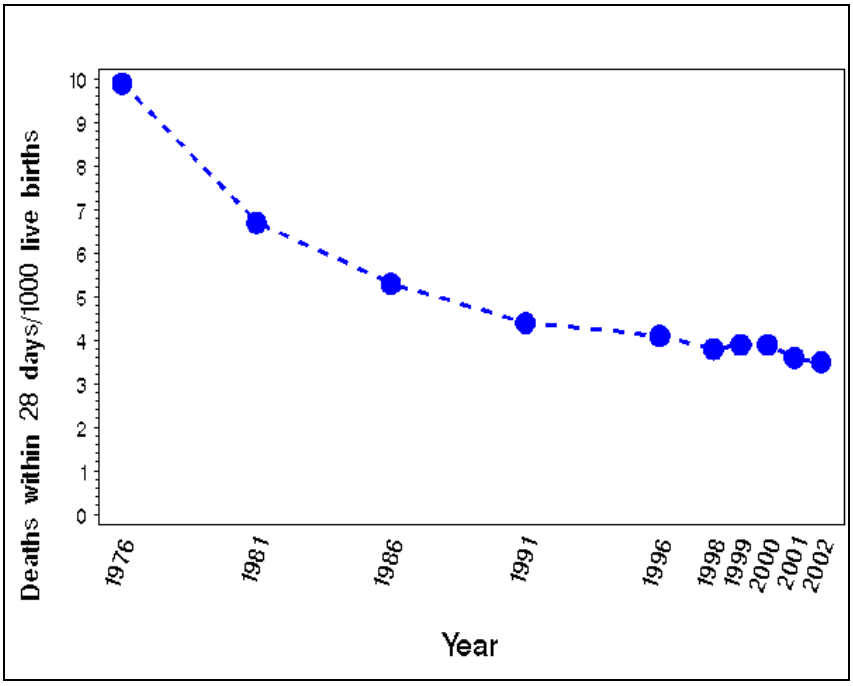
It is estimated that around 10% of babies born in the UK are admitted for neonatal care (Audit Commission, 1992). However, the majority of these infants are admitted for 'special care', for example jaundice requiring phototherapy, blood glucose monitoring (Rennie and Roberton, 2002:1), with about 2% of all babies requiring full intensive care.

It has been recommended that neonatal units should work together as Managed Clinical Networks (British Association of Perinatal Medicine, 2001:1), with at least one unit providing the full range of neonatal intensive care. Although this approach has not been implemented in the former Trent Health Authority area, there has been some work looking at the economic impact of implementing such networks (Draper *et al,* 2004). The units included in this thesis operate at level 2 or 3, since all babies born at 32 weeks gestational age or less, if surviving to admission, are likely to require intensive care or high dependency care at some stage of their admission. Hence, referral patterns, by medics or nurses, are unlikely to differ between units.

Over recent years there has been a fall in neonatal mortality in England and Wales (Office of National Statistics, 2003a); shown in Figure 2.3. A similar fall in mortality has been seen for

very low birth weight babies (< 1500g): around 50% mortality in 1975 falling to under 20% in 1995 (Tucker *et al*, 2004).

Figure 2.3 28-day Mortality by Year: England & Wales



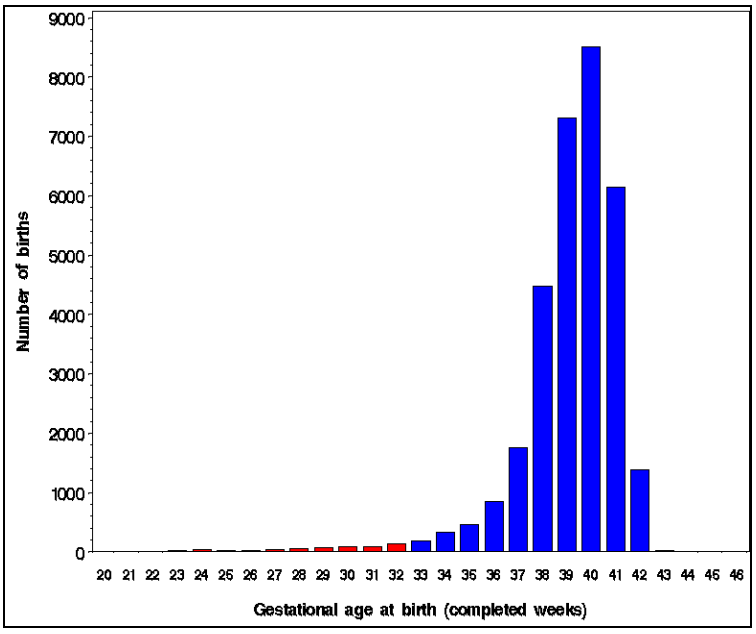
Improvements in care led to changes in registration practices (Wen *et al,* 2000), and to admissions to neonatal care, as infants previously considered not viable are now admitted. The changes in the characteristics of admitted infants can be complicated. While improvements in obstetric care have led to increasing numbers of very small infants, these improvements have also meant that such infants are in 'better clinical condition' when admitted (Baumer *et al,* 1997). However, there is some evidence that improvements in survival may not be consistent across all gestational ages. One study from British Columbia, Canada compared a cohort of infants born in 1983-1989 with those born in 1991-1993 (Battin *et al,* 1998). By the latter time period surfactant, antenatal steroids and dexamethasone were standard treatments for these infants. However, while survival rates had been seen to improve for 26-28 week infants, no improvement was seen for those born at 23-25 weeks gestational age.

Preterm birth

The aetiology of preterm birth is unclear, with about 80% of all preterm births resulting from spontaneous preterm labour, preterm rupture of the membranes or vaginal bleeding, and the other 20% attributable to obstetric intervention due to maternal or fetal indications (Goldenberg *et al,* 2000; Pschirrer and Monga, 2000). Many factors have been suggested

being related to preterm birth, including intrauterine infection (Goldenberg *et al,* 2000), low social class (Meis *et al,* 1995; Peacock *et al,* 1995), maternal stress (Dole *et al,* 2003), and, in sheep, a reduction in food intake around conception (Bloomfield *et al,* 2003). There is little evidence of the effectiveness of therapies proposed to prevent or arrest preterm labour, or that improvements in neonatal outcomes derive from improvements in perinatal and neonatal care (Goldenberg, 2002). Very preterm births, 32 weeks gestational age or less, only account for a small proportion of all births. Local data show that in the years 2000 to 2002 less than 2% of all births to Leicestershire resident mothers resulted in births at 32 weeks or less: 595 out of 32045 (Figure 2.4).

Figure 2.4 Births to Leicestershire resident mothers 2000-2002 by gestational age at birth



Quality of neonatal care

It is intuitive to hypothesise that the quality of care given by health care professionals can influence the outcome of infants. There is evidence to suggest that sub-optimal care can produce an increased incidence of poor outcomes. The Project 27/28 study investigated births at 26 to 29 completed weeks gestational age and identified associations between neonatal death and care given, including the timing of the administration of surfactant, the appropriate use of mechanical ventilation, early thermal care and the early use of inotropes (CESDI, 2003). There has been evidence shown that highly individualized care plans can reduce the amount of invasive treatment a very preterm neonate receives, although no difference was seen in mortality in this study (Fleisher *et al,* 1995).

Poor quality of care also comprises medical errors, and in the NHS it has been estimated that adverse events (errors) in which harm is caused to the patient occur in around 10% of admissions, over 850,000 per year (Expert Group on Learning from Adverse Events in the NHS, 2000). Neonatal medicine has particular problems with medical errors (Gray and Goldmann, 2004). A study in the USA found that although error rates in medication prescribing for neonatal intensive care units were similar to other types of wards, the rate of potential or preventable adverse drug events was much higher (Kaushal *et al,* 2001). Possible reasons suggested for this are that neonates are very often critically sick, neonatal weights change rapidly, neonates are less able to tolerate errors and that medicines usually do not come in doses suitable for neonates and need to be specially prepared. Often neonates require doses one tenth of those required by an adult and in these circumstances errors can lead to 10-fold overdoses (Chappell and Newman, 2003).

The evidence of an association between unit workload and neonatal mortality is equivocal. It has been hypothesized that outcome differences found between neonatal units in the UK and Australia were due to the more centralised care system in Australia producing better outcomes (International Neonatal Network, 2000). Some studies from the USA have shown evidence of reduced mortality rates in larger units (Phibbs *et al*, 1996; Rogowski *et al*, 2004), although at least one other study has not (Horbar *et al,* 1997). Such differences have previously been shown in the UK in 1988 to 1990 (The International Neonatal Network, 1993), other more recent work has shown no evidence of such a link (UK Neonatal Staffing Study Group, 2002). It has been suggested that the faster uptake of new therapies by tertiary (usually high volume) units compared to small units in the former time period may explain some of the differences between these groups of units (Parry *et al,* 2003a). Other work has found an opposite effect with better outcomes in the less centralised Danish neonatal intensive care provision compared to the former Trent Region (Field *et al,* 2002). However, varying case definitions and registration procedures often complicate such international comparisons.

Published TNS data from 1987 showed evidence that infants born at 28 weeks gestational age or less, and admitted to large units (defined at that time as >600 ventilator days per year), had better survival rates that those admitted to smaller units: 48% mortality versus 78% (Field *et al,* 1991). However, for infants born at 29 or 30 weeks gestational age there was a trend in the opposite direction (18% versus 7%), but this did not reach statistical significance at the 10% level ($p = 0.12$: Fisher's exact test). By 1994-1996 there was no evidence that such differences still existed (Field and Draper, 1999), and this change was thought to be due to

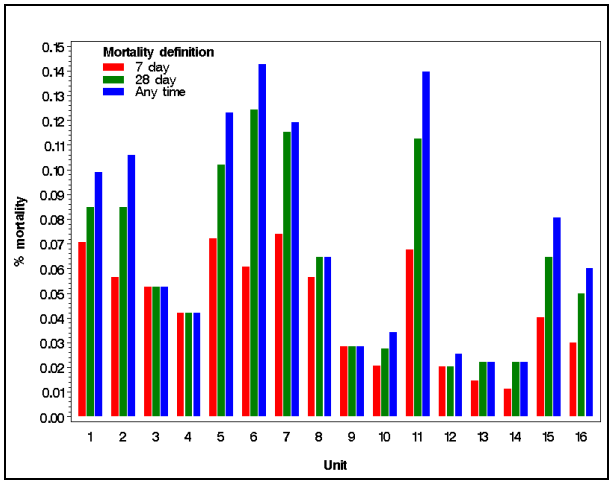
increased levels of specialist medical and nursing care and to the appropriate transferring of infants. However, there has been evidence presented of increased mortality with units working at high capacity (UK Neonatal Staffing Study Group, 2002).

Neonatal mortality

The outcome of interest in this thesis is death before discharge from the neonatal unit. There are several commonly used categories for the death of newborn infants: perinatal death (stillbirths and deaths under one week), neonatal death (under four weeks) and infant death (under one year). Neonatal death is sometimes further divided into early neonatal deaths (within seven days) and late neonatal deaths (seven to twenty-eight days). Such definitions have been used in previous studies, for example de Courcy-Wheeler *et al,* 1995, Zullini *et al,* 1997, Horbar *et al,* 1997.

The observed in-unit 7-day, 28-day and 'any time' death rates, for TNS data investigated in this thesis (introduced in §2.4), are shown in Figure 2.5. There were a total of 285 in-unit deaths, of which 160 (56.1% of all in-unit deaths) died within the first seven days of life and 244 (85.6%) died within the first 28 days. The differences in mortality rates, in particular between 7-day and 'any time' mortality, were greatest in the larger units.

Figure 2.5 7-day, 28-day and any time in-unit mortality by neonatal unit



In this thesis total in-unit mortality was investigated rather than death within a given time. There were four reasons for this choice:

i) Some infants may die a short time after discharge and TNS does not collect information about such deaths; although the number of infants discharged home where it is felt that the infant was likely to die as a result of their current morbidity is very low. However,

there is local anecdotal evidence of at least one infant being discharged home for palliative care (Field, D.J.: Personal communication);

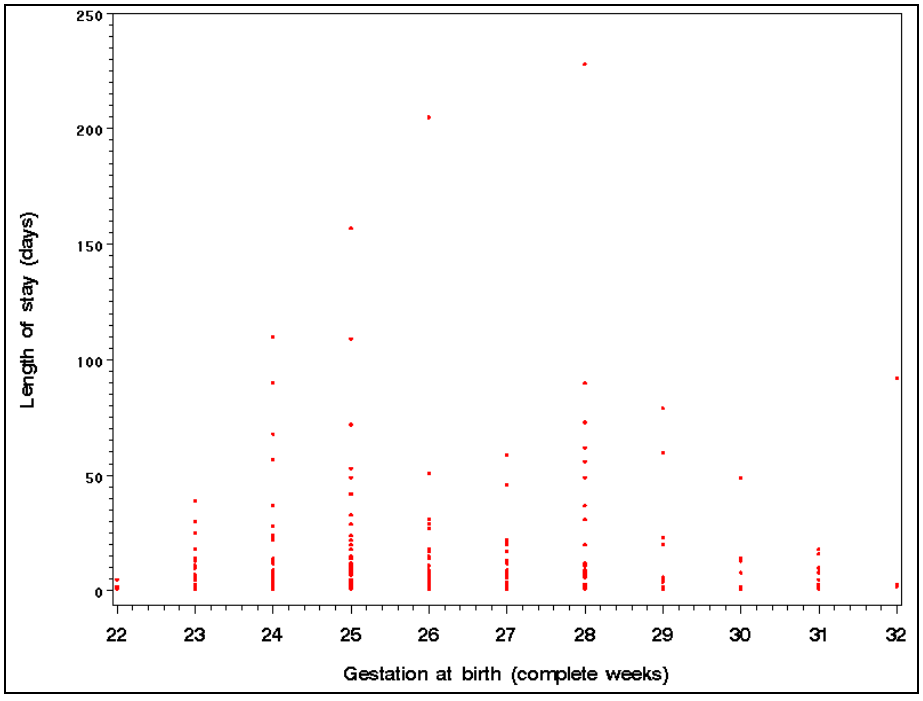
ii) The aim is to use mortality rates as a measure of care received within the neonatal units. Once the infant has been discharged from the unit it can be argued that subsequent death may be attributable to factors outside the control of the unit. Inappropriate discharge, where the infants went on to died due to existing morbidity, is likely to be extremely rare;

iii) Any choice of a cut-off point would be arbitrary and may not be appropriate over the whole range of gestations at birth;

iv) Some deaths may occur on the neonatal unit after the cut-off, but any death on a neonatal unit is important.

In effect, given the low probability of death after discharge, the choice of in-unit mortality could be seen as a close approximation to death within 228 days of birth (the longest time between birth and in-unit death observed in the data used in this thesis: Figure 2.6).

Figure 2.6 Length of stay for infants who died before discharge



Some studies have investigated subsequent survival for infants who have survived to a certain time, for example the first three days of life (Fowlie *et al,* 1998). While such an approach may be useful for informing parents, allocating infants to strata within a clinical trial and allocating

care, the fact that many deaths would be excluded means that it is an unsuitable approach for profiling here.

It is acknowledged that there is disagreement about the use of mortality as a measure of care. In particular, the increasing survival rate of very low birth weight infants means that the chronic morbidity (particularly neurological) experienced by these infants is also very important (McCormick, 1997; Bard, 1993; Colver *et al,* 2000; Allen, 2002). However, mortality has been chosen here as it has been reliably collected in the Trent Neonatal Survey and it is, in itself, a very important outcome and of interest to neonatologists and administrators within the Region (The Trent Infant Mortality and Morbidity Studies, 2003). As discussed in §2.3, there is also some evidence that mortality rates are influenced by the type of care given. If true, mortality may be an appropriate measure of the quality of care provided.

In-unit survival is a commonly used outcome to compare neonatal units. However, other outcomes have been used in published papers. Examples of outcomes used include: alive without supplementary oxygen on day 28 (Horbar *et al,* 1988); rates of chronic lung disease (CLD) at 36 weeks adjusted gestational age, retinopathy of prematurity (ROP), intraventricular haemorrhage (IVH), patent ductus arteriosus (PDA), nosocomial infection, stage 2 or greater necrotising enterocolitis (NEC), survival without major morbidity (grade 4 IVH, CLD NEC or grade 4 ROP) (Lee *et al,* 2000), blood transfusion rates (Bednarek *et al,* 1998) and narcotic use (Kahn *et al,* 1998). However, these outcomes are either dependent on unit clinical policy and unsuitable for inferring the level of care of individual units (e.g. alive without supplementary oxygen and rates of CLD are hugely influenced by policy on oxygenation) or are outcomes not recorded by TNS (e.g. ROP, NEC and PDA).

2.4 Study Population and Outcome

The source of the data in this thesis (the Trent Neonatal Survey) was described in §2.2. The data to be analysed are introduced in this Section, together with the problems of implausible values for birth weight for gestational age and how infants who may have received care from multiple units will be allocated to an individual unit.

Study population

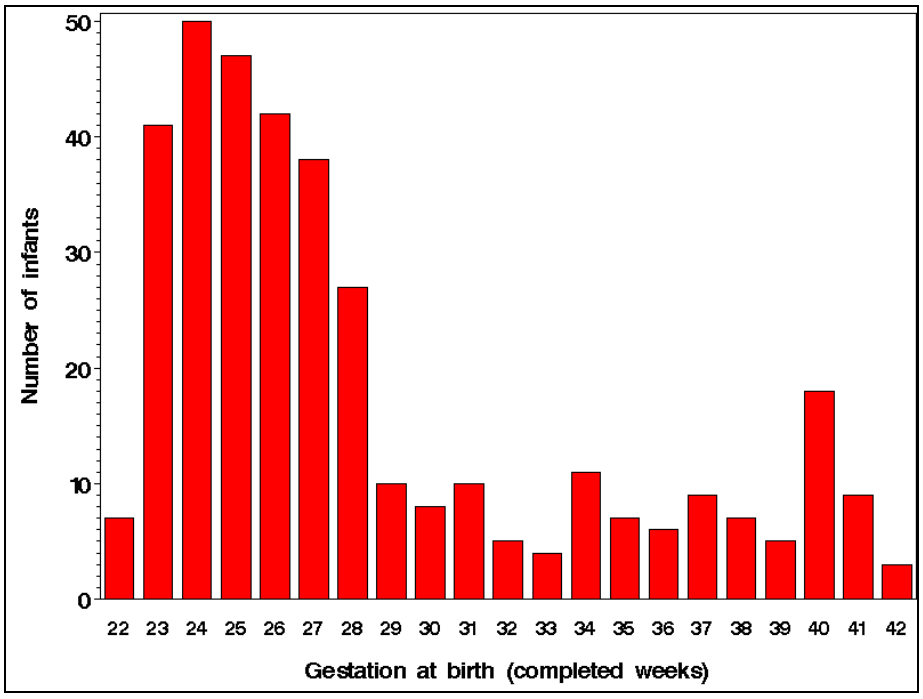
With any study it is important to carefully define the population of interest. The choice of population of preterm infants can influence study results (Evans and Levene, 2001): for example should labour unit deaths be included (i.e. all live births) and what about stillbirths, both antepartum and intrapartum? This thesis has only looked at infants admitted to neonatal care. However, this could exclude some infants who could not be, or were not, resuscitated on the labour unit, a decision likely to have been made by the neonatal team. Other studies, when considering preterm births, have looked at all deliveries within a hospital (Horbar *et al,* 1997), whether they survived to admission or not. However, in this thesis interest was with the care provided by the neonatal units and, therefore, only admissions were considered. Of course, there are likely to be differences in the types of babies admitted to individual units, especially at the extremes. There is a debate on what type of care should be given to infants, particularly very preterm infants, even if their chances of survival are felt to be negligible (Greisen, 2004; Levene, 2004; Morrison and Rennie, 1997). It may be argued that good nursing care is more appropriate than subjecting the infant to invasive intensive care therapies with little hope of survival. It is likely that different clinicians (together with parents and nursing staff) will have different opinions, and that hospitals will have different policies. An Australian study showed that out of 71 neonatologists, 77% said that they would never resuscitate a 22-week infant, and 11% said that they would never resuscitate a 23-week infant (Oei *et al,* 2000). In another Australian study, a sample of experienced obstetricians from non-tertiary hospitals stated that the minimum gestational age at birth at which they would consider active intervention ranged from 22 to 26 weeks (Gooi *et al,* 2003). In the Netherlands active intensive care is generally not given to infants born at less than 25 or 26 weeks gestational age (Sheldon, 2001). At the extreme, there is anecdotal evidence of infants born alive following failed abortions not being given treatment in some hospitals but being admitted to neonatal units in others, with reports of two such cases being admitted to Leicester Royal Infirmary (Templeton and Rogers, 2004). It is clear that infants at the margins of viability will have very poor prognoses. A neonatal unit that takes an aggressive interventionist policy with such infants is very likely to show high death rates, not necessarily due to poor quality of care but, rather, because of its admission policy. Certainly, units which admit a large number of infants with very poor prognoses run the risk, if model risk adjustment is not perfect, of having increased observed mortality over what is expected. The TNS does not collect information on infants not admitted but, since the neonatologists are

making decisions on admission, it should at least be acknowledged that there is some information on performance being lost.

In some studies, for example Lee *et al* (2000), all admissions have been considered, no matter the weight or gestational age at birth. The data available to this thesis only allow a subgroup of admissions to be considered, as data are not collected on all admissions to neonatal intensive care units (see the inclusion criteria in §2.2). However, the majority of neonatal deaths occur to preterm infants (Rennie and Roberton, 2002:1). Indeed, local information confirms that infants born at 32 completed weeks achieve 98% to 99% survival (Draper *et al,* 2003). It has been suggested that the greater heterogeneity of the pathology of term babies and the greater unpredictability of their outcome make outcomes in that group a more sensitive marker for quality of care (Marlow, 2002). However, the unpredictability of death in these more mature infants, and the small number of deaths, would make it difficult to use them to measure the quality of care.

The total number of deaths, of all gestational ages at birth, recorded on neonatal units within the Trent Region from 2000 to 2002 was 364, with 285 (78%) of these infants born at less than 33 complete weeks gestational age (Figure 2.7). Thus, the data used in this thesis include the majority of in-unit deaths, but not all. However, the pathology of term infants dying is likely to be different to preterm births so this distinction may be appropriate.

Figure 2.7 Total observed in-unit mortality by gestational age at birth: TNS data for all gestational ages



A further point of note is that this thesis investigates births at 32 weeks gestational age or less: the cohort is not defined by birth weight. Some previous studies have used a combined selection criteria, such as less than 32 weeks or less than 1500g birth weight, in an attempt to include all 'small' infants, for example de Courcy-Wheeler *et al* (1995). The TNS uses both of these as criteria for inclusion into the survey, although clinical investigations usually use subsets defined by gestational age, for example Draper *et al* (1999), Lal *et al* (2003), Manktelow *et al* (2001). The use of birth weight to define a cohort potentially introduces bias into an analysis because a single definition of 'low birth weight' is unlikely to be appropriate for all group within a population (Wilcox, 2001).

The population selected from TNS to be investigated in this thesis were all infants admitted to the sixteen NICUs named above, who were deemed to have been born at less than 33 weeks gestational age in the years 2001, 2002 and 2003. Infants with lethal congenital abnormalities were excluded. There were 3063 infants identified and, of these, 37 (1.2%) were identified with lethal abnormalities and were excluded from all further analyses, leaving 3026 infants.

2.4.1 Plausible birth weight for gestational age

The data on the 3026 remaining infants were inspected for recorded weight and gestational age at birth. There are several difficulties in estimating gestational age, discussed in more detail in §6.4.1. This, together with the potential for recording errors in both gestational age and birth weight, means that it is useful to check the data for possible implausible values of birth weight for gestational age. An exceptionally high, or low, birth weight at any given gestational age may indicate an error in the data. This is more likely, unless it is the result solely of a recording error, to be an error in the recorded gestational age than the birth weight. The incorrect inclusion of infants actually born at greater than 32 weeks gestational age is likely to result in a falsely low estimated mortality rate since infants born later have higher rates of survival.

Sometimes such observations thought to be implausible have been removed after simply inspecting the data, for example Gruenwald (1966), but in other cases various rules for identifying implausible values have been proposed. Some of these are simple rules, such as birth weights more that two interquartile ranges above the 75th, or below the 25th, birth weight for gestational age centile (Arbuckle *et al,* 1993), more than two standard deviations above the mean weight for gestational age (Seeds and Peng, 1998), more than 40% over the mean birth weight for gestational age (Arnold *et al,* 1991), over six standard deviations from the

mean (Cheung *et al,* 2000) and four or five standard deviations above or below the median birth weight for gestational age (Joseph *et al,* 2001).

Other methods use the assumption that the distribution of birth weights conditional on gestational age follows a Normal distribution and then delete or reassign gestational age to extreme observations (Zhang and Bowes, 1995). A more sophisticated method uses a mixture model to estimate the conditional birth weight distribution assuming a conditional Normal distribution (Platt *et al,* 2001). It is further assumed that errors in recorded gestational age at birth resulted in either the gestational age being correctly recorded or a term birth incorrectly recorded. In their paper, the assumption made by Platt *et al* was that all preterm births were correctly recorded. Other error patterns were also referred to, but not presented as the solutions to such models are computationally very intensive. A further approach has been suggested, similar to Platt *et al* but this time assuming a conditional log-normal distribution for the birth weights and also assuming that all errors are ± 4 weeks. Extreme observations of birth weight for gestational age can then be identified using local criteria (Oja *et al,* 1991).

When the empirical distribution of birth weight at each gestational age is multimodal, one proposed method is to assume that the lowest observed mode is the true mode for the distribution. If it is further assumed that the true distribution is symmetrical about the mode, and that the values below this observed mode are correct, percentiles can then be estimated and observations above a specified percentile can be identified (David, 1983)

An alternative approach is to investigate the range of gestational age for a given birth weight since, it is argued, birth weight can be measured more accurately than gestational age. Alexander *et al* (1996), using observed data from over 3,000,000 singleton live births in the USA in 1991, divided the observations into groups according to their birth weight (125g intervals) and then investigated any observations with reported gestational ages more than 2.5 standard deviations from the mean gestational age for that birth weight group. Such outlying observations were then further inspected for possible deletion using clinical judgement.

While most of the methods outlined above assume a unimodal distribution for birth weight conditional on gestational age, preterm births are, by definition, a select group of fetuses (i.e. fetuses delivered prematurely) and a multimodal conditional distribution may, potentially, be more appropriate. The distribution has been modelled at least once before as a mixture of two Normal distributions (Milner and Richards, 1974). The observed conditional distribution for the data in this thesis is investigated in Appendix G.

Such rules only identify possible implausible birth weights and not other possible errors in the reported gestational age or birth weight. It is quite possible for an observation with an incorrectly reported gestational age to have a birth weight that appears quite plausible. No matter which cut-offs are chosen it is inevitable that some incorrect values will remain, just as some true but extreme values may have been excluded (Platt *et al,* 2001; Altman and Chitty, 1997). Indeed, it is only gross errors that will be identified using these methods. There has been evidence presented that those excluded by such rules tend to have higher birth weight specific and gestational age specific mortality rates than other live births (Joseph *et al,* 2001; Parker and Schoendorf, 2002). Nevertheless, these are useful techniques for identifying problems.

The data in this thesis show no evidence of multimodal conditional birth weight for gestational age distributions (§6.4.3), characteristic of term births incorrectly recorded as preterm (Skjaerven *et al,* 2000; Platt *et al,* 2001; Altman and Chitty, 1997; Parker and Schoendorf, 2002). However, this is not surprising as infants in these data are included only if they were admitted to a NICU, which is likely to have resulted in any large errors in the recording of gestational age to have been noticed and corrected. The TNS already substitutes gestational age estimated from early ultrasound scan when the last menstrual period dates are thought to be incorrect and, therefore, already uses some form of data modification. Any extreme errors remaining are likely to be recording errors. Because of this, together with the tendency for such methods to exclude high-risk infants, this thesis will use the method proposed by Alexander *et al* (1996). This method gives wide, conservative limits (Parker and Schoendorf, 2002) as well as being readily available and easy to use. These limits are given in Table 2.1 and are shown in Figure 2.8 together with the observed birth weights.

Only one observation fell outside the limits proposed: 2915g and 29 weeks gestational age at birth (Figure 2.8). After confirmation of these values on the original TNS questionnaire, this observation was excluded and the remaining 3025 observations were used for all further analyses, unless otherwise stated.

Figure 2.8 Birth weight by gestational age: TNS data

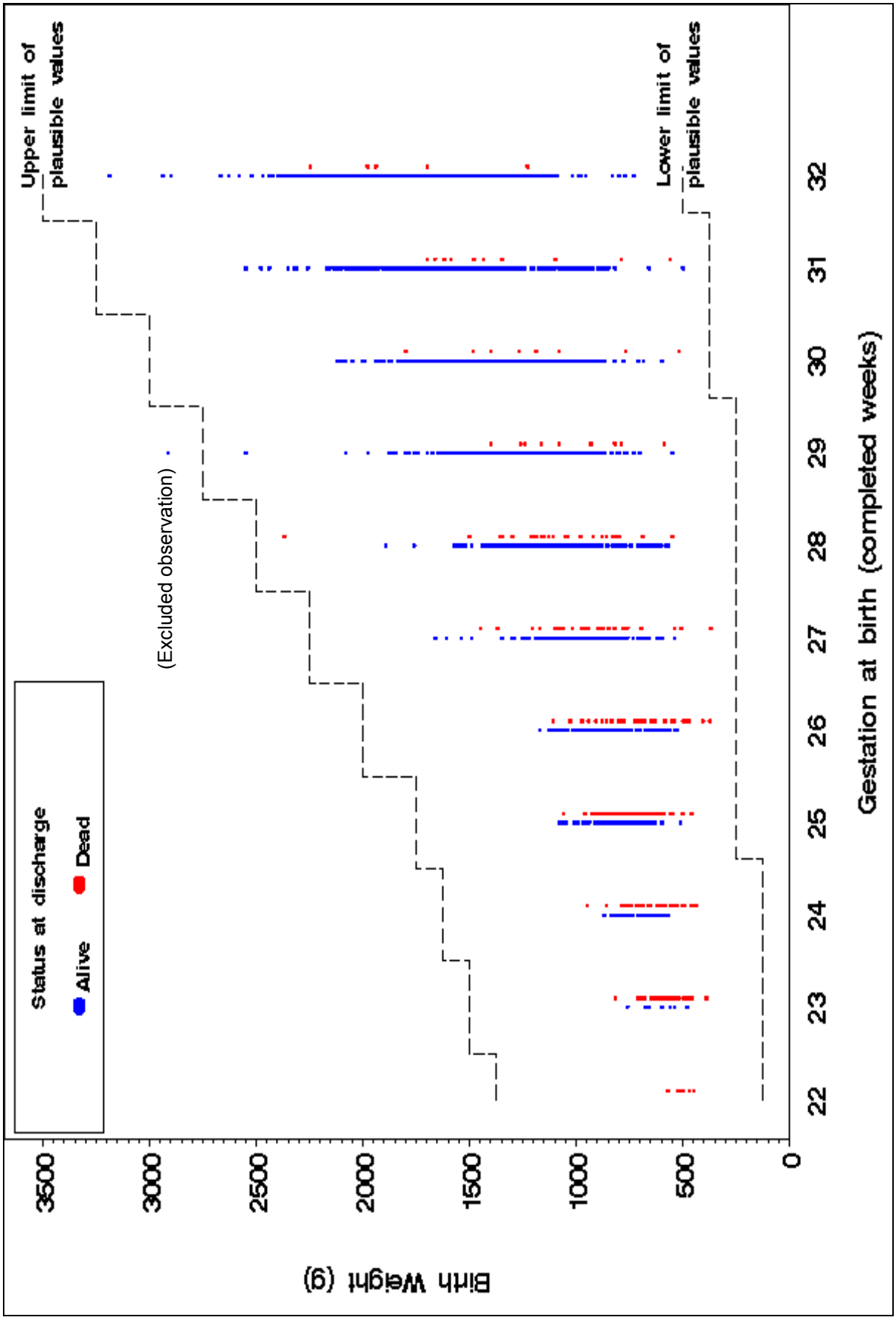


Table 2.1 Limits of plausible birth weight

Gestational age (weeks)	Plausible birth weight[f] (g)	
	Lower limit	Upper limit
22	125	1375
23	125	1500
24	125	1625
25	250	1750
26	250	2000
27	250	2250
28	250	2500
29	250	2750
30	375	3000
31	375	3250
32	500	3500

2.4.2 Allocation to NICU

One notable feature of neonatal care in the UK is that many infants are transferred between neonatal intensive care units. This may be before birth (in-utero) if the anticipated level of care and medical expertise, either for the birth or postnatally, is not available in the booked hospital of birth. Transfers also occur postnatally if the level of care required is greater than that which can be given at the hospital of birth. It is also usual for an infant to be transferred back from a referral centre to its local unit once it is possible to do so. While these types of transfers are appropriate to fulfil clinical requirements, other transfers may not be. In particular, transfers may occur because a unit does not have sufficient capacity at that time. A 1999 census of the 37 largest perinatal centres in the UK showed that a proportion of postnatal transfers out occurred due to lack of cots or appropriate staff (Parmanum *et al,* 2000). The 2002 TNS Annual Report reported that 181 infants recorded by the survey in that year were inappropriately transferred (The Trent Infant Mortality and Morbidity Studies, 2003) although the reasons were not known. Currently, transfers are generally carried out in a unplanned manner with little formal organisation between units (Field *et al,* 1997; Fenton *et al,* 2004).

[f] From (Alexander *et al*, 1996)

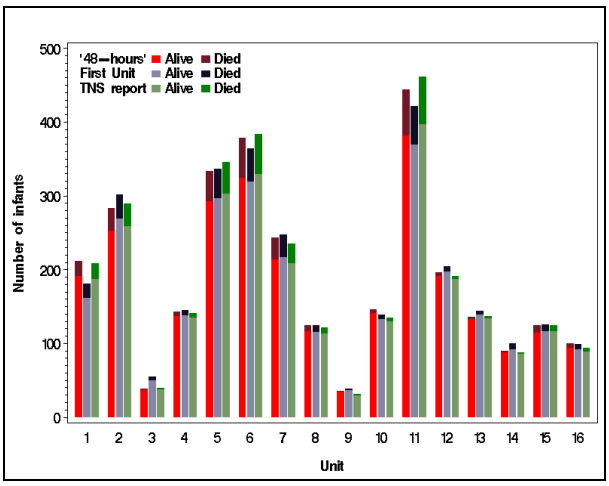
For the purposes of analyses, infants need to be allocated to a single unit. One method is to attribute the unit that provided the longest period of care between 12 and 72 hours after birth (Scottish Neonatal Consultants' Collaborative Study Group and the International Neonatal Network, 1995; Parry, 1998). However, the Trent Neonatal Survey only records that an infant was on a particular unit on a given day and, thus, does not provide sufficient detail to use this method of allocation. The decision was therefore made that when an infant has been transferred, the NICU of care should be the first where the infant was recorded as having been on the unit on at least three consecutive days. It is acknowledged that, depending on the time of admission and discharge, the minimum time on the unit may range from 24 to 72 hours. If any transferred infant did not stay on any unit for this minimum amount of time, it was assigned to a unit using clinical judgement, considering the time spent on the units, the care given and the type of unit.

Of the 3025 infants included in this study, 2814 (93%) were assigned to their first unit of admission and 2455 (81%) were never transferred to another unit. Of the 570 who were transferred, 359 (63%) were assigned to the unit of their first admission and 204 (36%) were assigned to their second unit. One infant was assigned to the third unit of admission having spent less than two days in two different units before being admitted to a third unit where he stayed for 12 days before being discharged home. In addition, there were six infants who were transferred between units, but who were not on a single unit for three consecutive calendar days. All of these infants died before discharge. These infants were allocated to their second unit of admission. This choice was made for three reasons. For each infant the stay on the second unit (two days) is as least as long as the first unit (one or two days), the second unit is a referral unit implying that the transfer was appropriate and the deaths occurred on the second unit.

The sensitivity of the numbers to allocation procedure can be investigated by using different criteria. In Figure 2.9 the number of admissions, by outcome, are shown using three different allocation methods: '**48-hours**' represents the method used in this thesis by allocating an infant to the first unit in which it stayed for three consecutive days; for '**First Unit**', the infant was allocated to the unit of first admission regardless of the length of stay; '**TNS report**' reproduced the allocation method used for the TNS report by allocating infants to the second unit of admission in the case of emergency (flying-squad) transfers but to the first unit otherwise.

Although the methods resulted in different numbers of infants for each unit, the variation was small. The decision to allocate infants to the first unit in which they stayed for three consecutive days was preferred as it matched infants to the unit where most early care was received.

Figure 2.9 Admissions and deaths in units by method of allocation



2.4.3 Infant characteristics

Important characteristics of the TNS data are described below: more detailed investigations are left until Chapter 6.

The distribution of reported gestational ages is shown in Figure 2.10. Unsurprisingly, there are increasing numbers of admitted infants with increasing gestational age. This is mostly due to more births at higher gestational ages (in the range investigated here) but also due the increased likelihood of survival to admission to a NICU. The distribution of recorded birth weight is shown in Figure 2.11. However, it is birth weight for gestational age that is likely to be more informative and this is discussed in more detail in §6.4.3.

Figure 2.10 Histogram of observed gestational age at birth for TNS data

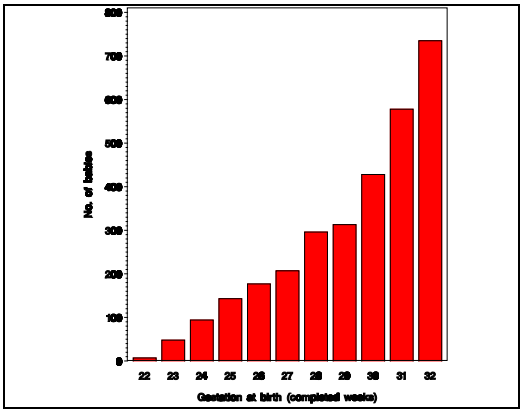
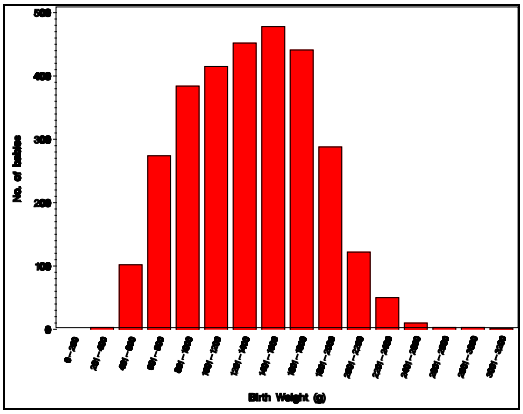


Figure 2.11 Histogram of observed weight at birth for TNS data



More than half of the infants admitted were boys (Table 2.2). The one infant for whom sex was not recorded was born at 30 weeks gestational age and was discharged home alive. Although the majority of infants were singletons, around one quarter were from multiple pregnancies (Table 2.3). The ethnic group of the infant, as reported by the mother, is recorded by TNS (Table 2.4).

Table 2.2 Observed number of infants by sex of infants for TNS data

Sex	No.	(%)
Male	1667	(55.1)
Female	1357	(44.9)
Not recorded	1	(0.0)
Total	3025	

Table 2.3 Observed number of infants by multiplicity of pregnancy for TNS data

Multiplicity of pregnancy	No.	(%)
Singleton	2288	(75.6)
Twin	669	(22.1)
Triplet	68	(2.2)

Table 2.4 Observed number of infants by ethnic group for TNS data

Ethnic group of infant	No.	(%)
European	2551	(84.3)
Asian	239	(7.9)
African / West Indian	52	(1.7)
Mixed race	108	(3.6)
Other	14	(0.5)
Not known / not recorded	61	(2.0)

The observed mortality before discharge for each unit is given in Table 2.5. There is a wide variation between the units in the proportion of infants who died before discharge: from 2.2%, in units 13 and 14, to 14.3% in unit 6.

Table 2.5 Observed In-Unit Mortality 2000-2002 for TNS data

Unit	Total Infants	No. Died	Percentage of Admissions	(Exact 95% confidence interval)
1	212	21	9.9	(6.2 to 14.8)
2	283	30	10.6	(7.2 to 14.8)
3	38	2	5.3	(0.6 to 17.8)
4	142	6	4.2	(1.5 to 9.0)
5	333	41	12.3	(8.9 to 16.4)
6	378	54	14.3	(10.9 to 18.3)
7	243	29	11.9	(8.1 to 16.7)
8	124	8	6.4	(2.8 to 12.4)
9	35	1	2.9	(0.0 to 15.0)
10	146	5	3.4	(1.1 to 7.9)
11	445	62	13.9	(10.8 to 17.5)
12	196	5	2.6	(0.8 to 5.9)
13	136	3	2.2	(0.4 to 6.4)
14	90	2	2.2	(0.2 to 7.8)
15	124	10	8.1	(3.9 to 14.4)
16	100	6	6.0	(2.2 to 12.6)
Total	**3025**	**285**	**9.4**	**(8.4 to 10.6)**

Exact 95% confidence intervals are also shown in Table 2.5, calculated using a link between the binomial and the *F* distributions (Armitage and Berry, 1994:121). The limits for the interval are given by:

$$\pi_{Lower} = \frac{r}{r + (n - r + 1)F_{0.025,2n-2r+2,2r}} \qquad (2.1)$$

$$\pi_{Upper} = \frac{r+1}{r + 1 + (n - r)F^{-1}_{0.025,2r+2,2n-2r}} \qquad (2.2)$$

where: n is the number of observations

r is the number of events $\quad r = \sum_{i=1}^{n} d_i$

2.5 Chapter Summary

In this Chapter the data source (the Trent Neonatal Survey) and the data investigated in this thesis were introduced. Some characteristics of neonatal intensive care and its organization were also described.

In §2.4 basic descriptive statistics for the data were reported and the problem of implausible reported birth weight for gestational age was discussed, resulting in one observation being excluded from further analyses. The method of allocating infants to responsible NICUs was also described.

Observed in-unit mortality rates for the NICUs were reported in Table 2.5 and a wide variation in rates noted. The observed differences can be accounted for by three sources of variation (as discussed in §1.3.2): random variation, differences in case-mix and differences in the type of care between the units. Statistical methodology to disentangle these causes will be described in the rest of the thesis, beginning with methods to quantify the random variation for provider profiling with binary indicators in Chapter 3.

Chapter 3: STATISTICAL METHODOLOGY

3.1 Chapter Overview

In the previous Chapter, Table 2.5 showed the observed in-unit mortality rates for the neonatal units in this thesis. The differences in rates observed can be accounted for by three sources: random variation, differences in case-mix and differences in the type of care between the units (§1.3.2). In this Chapter statistical methods that quantify the random variation, and have been proposed for provider profiling with binary indicators, will be introduced and discussed. Case-mix adjustment is taken up in Chapter 4.

Section 3.2 comprises a brief description of two approaches to statistical analysis used in this thesis: Classical and Bayesian. In §3.3 simple statistical methods for provider profiling using binary indicators are described and illustrated. It is shown that such methods are unlikely to be sufficiently detailed or flexible to allow robust conclusions to be drawn and a statistical modelling approach is suggested as more appropriate. This leads to §3.4 where generalized linear models, and in particular logistic regression models, are introduced and illustrated. Other statistical methods that may be applicable to provider profiling in general, but not suitable for the data in this thesis, are briefly described in §3.5. The main results and conclusions from the Chapter are reported in §3.6.

3.2 Frequentist and Bayesian statistical methods

The statistical methods discussed in this thesis can generally be undertaken using either **Frequentist** (**Classical**) or **Bayesian** methods. Although the primary analyses in this thesis will use frequentist methodology, a Bayesian approach will sometimes be presented to illustrate this potential approach and the results obtained compared with those from the frequentist analysis. The difference between the two approaches is only briefly discussed, as the emphasis is on contrasting the different interpretations of the results obtained. However, it is felt that a brief description of the main characteristics of each approach may be useful.

3.2.1 Frequentist methods

Frequentist statistical methods are based on the **relative frequency** concept; i.e. the proportion of times an event occurs over a number of observations. The long-term frequency of an event occurring (the limiting relative frequency) is interpreted as the probability of that event occurring on a single occasion (Lindley, 2005).

Much of classical statistical methods involves **hypothesis testing** and the calculation of confidence intervals. Hypothesis testing involves specifying a hypothesis of interest: the **null hypothesis** (H_0). In the case of a two arm clinical trial, the null hypothesis is often that the population mean value of the response of interest is the same in each group: i.e. $\mu_A = \mu_B$, where A and B signify the treatment groups and μ the population mean. This hypothesis of no difference is sometimes referred to as the **nil hypothesis** (Cohen, 1994), which may help as a reminder that null hypotheses does not solely have to be of the form of no difference. Instead, the null hypothesis could, and some say should (Mulaik *et al,* 1997:68), be used to test a hypothesised difference between the two groups, e.g. $\mu_A = (\mu_B + 4)$.

The approach involves the calculation of the probability of obtaining the study results, or results more extreme, if the null hypothesis was true, i.e. $\boldsymbol{P(data|H_0)}$**,** where ***data*** is the totality of data equal to or more extreme than that observed (Berry and Stangl, 1996). A small value of P is taken as evidence that the data are unlikely to have come from a population with the hypothesised parameter values and, hence, it is argued, the null hypothesis can be rejected.

The uncertainty around the estimate of a parameter or statistic is usually presented using a 100(1-α)% confidence interval. Such intervals have the interpretation that over repeated sampling the true value lies in such intervals 100(1-α)% of the time.

3.2.2 Bayesian methods

It may be argued that it is the probability of a hypothesis given the data, i.e. $\boldsymbol{P(H_0|data)}$, which is really the probability of interest. An alternative way of looking at this is to consider $P(\theta|data)$, the conditional probability of the parameters. However, $P(\theta|data)$ makes a probability statement about the parameters and in classical statistics parameters have fixed, but perhaps unknown, values. Using this approach requires the acceptance that a parameter can have probability distribution rather than being a constant. Once this is accepted $P(\theta|data)$ can be calculated using **Bayes Theorem**:

$$P(\boldsymbol{\theta} \mid data) = \frac{P(data \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(data)} \qquad (3.1)$$

where: θ is any parameter, for example $\theta = (\mu_A - \mu_B)$.

It can been seen that the use of Bayes Theorem requires $P(\boldsymbol{\theta})$, the unconditional density of $\boldsymbol{\theta}$, to be stated. This is the prior state of knowledge, or belief, about the hypothesis, or parameter, and is called the **prior probability distribution**. The formal inclusion of prior knowledge into the calculations is perhaps the most contentious aspect of the Bayesian approach. To some Bayesians this is the most important distinction between the two approaches (Berry and Stangl, 1996:8) as all evidence can be formally included in the analysis. Frequentists, however, may argue that this introduces subjectivity, since different people can draw different conclusions from the same data (Fisher, 1996).

In (*3.1*) $P(data|\boldsymbol{\theta})$ is the likelihood function for the data evaluated at $\boldsymbol{\theta}$, sometimes denoted as $\boldsymbol{L(\theta|data)}$. Since $P(data)$ is a normalising factor, then (*3.1*) can be rewritten:

$$P(\boldsymbol{\theta} \mid data) \propto L(\boldsymbol{\theta} \mid data)P(\boldsymbol{\theta}) \qquad (3.2)$$

Then, if all the parameters $\boldsymbol{\theta} = \theta_1, ..., \theta_k$ are assumed to be independent then:

$$P(\boldsymbol{\theta} \mid data) \propto L(\boldsymbol{\theta} \mid data)P(\theta_1)...P(\theta_k)$$

One method to obtain the marginal posterior distributions of each parameter is to integrate out the other parameters. For example, this gives for θ_1:

$$P(\theta_1 \mid data) \propto \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} P(\boldsymbol{\theta} \mid data)\, d\theta_2 ... d\theta_k \qquad (3.3)$$

Generally, the solution to equations such as (*3.3*) is not able to be obtained analytically and other methods are often used (Gefland, 1995). These include numerical evaluation, analytic approximation (e.g. Laplace approximation) and Monte Carlo integration (including Markov Chain Monte Carlo). The method to be used in this thesis is Markov Chain Monte Carlo (MCMC), and in particular Gibbs sampling. This method is felt to be the most appropriate for models with a large number of parameters to be estimated (Gilks *et al,* 1993).

Gibbs sampling

Characteristics of a posterior density function (e.g. mean, variance) can be expressed as functions of $\boldsymbol{\theta}$. The posterior expectation of such a function is given by:

$$E[f(\theta)|\,data] \propto \int f(\theta)L(\theta\,|\,data)P(\theta)d\theta \qquad (3.4)$$

Monte Carlo intergration evaluates $E[f(\theta)|\,data]$ by sampling $\{\boldsymbol{\theta}_t,\ t = 1,\ldots,n\}$ from $L(\boldsymbol{\theta}\,|\,data)P(\boldsymbol{\theta})$ and the ergodic averaging gives (Gilks *et al*, 1995:4):

$$E[f(\boldsymbol{\theta})|\,data] \approx \frac{1}{n}\sum_{t=1}^{n} f(\boldsymbol{\theta}_t) \qquad (3.5)$$

In practice it is often impossible to sample directly from $L(\boldsymbol{\theta}\,|\,data)P(\boldsymbol{\theta})$ as such distributions can be very non-standard (Spiegelhalter *et al*, 2004:105). One solution is to construct a Markov chain with the same state space as (*3.2*) and with equilibrium distribution $L(\boldsymbol{\theta}\,|\,data)P(\boldsymbol{\theta})$. Sampling from the chain, once the equilibrium distribution has been reached, will provide observations that can be used to estimate the required summary statistic of the posterior distribution, i.e. (*3.5*). Although this will not be an independent sample, since for a Markov chain any observation x_p is a function of x_{p-1}, convergence to the expectation of interest will be achieved if the chain is run for a sufficiently long time (Gilks et al, 1995:5).

The Gibbs sampler can produce a Markov chain with the properties outlined above (Smith and Roberts, 1993). If the joint posterior distribution is given by $P(\boldsymbol{\theta})=P(\theta_1, \theta_2,\ldots, \theta_p)$, then let $P(\theta_j|\ \boldsymbol{\theta}_{(-j)}, data)$ represent the induced full conditional distribution of parameter j given the value of the other parameters. The Gibbs sampler starts with initial values for the parameters $(\boldsymbol{\theta}^0)=(\theta_1^0, \theta_2^0,\ldots, \theta_p^0)$, then successive random observations are made from the full conditional distributions $P(\theta_j|\ \boldsymbol{\theta}_{(-j)}), j = 1,\ldots, p$. The sampling starts thus:

θ_1^1 from $P(\theta_1|\ \theta_2^0,\ldots, \theta_p^0, data)$

θ_2^1 from $P(\theta_2|\ \theta_1^1, \theta_3^0,\ldots, \theta_p^0, data)$

..

θ_p^1 from $P(\theta_p|\ \theta_1^1,\ldots, \theta_{p-1}^1, data)$

Hence, $(\boldsymbol{\theta}^0)=(\theta_1^0, \theta_2^0,\ldots, \theta_p^0)$ has been changed to $(\boldsymbol{\theta}^1)=(\theta_1^1, \theta_2^1,\ldots, \theta_p^1)$. Repeated application of the algorithm (say m times) produces a series of observations $\boldsymbol{\theta}^0$, $\boldsymbol{\theta}^1,\ldots, \boldsymbol{\theta}^m$. These are realisations from a Markov chain with an equilibrium distribution equivalent to the joint posterior distribution.

However, the Gibbs algorithm will not always produce a sample from the full sample space of the target posterior distribution. For example, a high level of autocorrelation within the chain can result in poor mixing and the chain not covering the full sample space. Methods exist to

try to improve the mixing of the chains and to diagnose any problems that may have occurred. Some of these are discussed below.

Credible intervals

The Bayesian equivalent to confidence intervals are credible intervals. However, since it is now assumed that parameters have probability distributions, intervals can be constructed that truly have the interpretation that there is a probability of (1-α) that the true parameter estimate lies within a 100(1-α)% credible interval. The intervals used in this thesis are two-sided **equi-tail-area** intervals. The interval limits (θ_L, θ_U) are defined where $P(\theta < \theta_L) = 0.025$ and $P(\theta > \theta_U) = 0.975$. However, for skewed distributions some values outside the interval may have higher posterior probabilities than some values in the interval. An alternative approach, to overcome this problem, would have been to use **highest posterior density** (HPD) intervals. In this case the values of the limits are selected so that they have identical posterior probabilities (Spiegelhalter *et al*, 2004:65). However, the limits of HPD credible intervals can be difficult to obtain, so equi-tail-area intervals were used in this thesis.

Application of Gibbs sampling to data

In this thesis, Gibbs sampling methods were applied to the data using the WinBUGS software (Spiegelhalter *et al*, 1999b). To start the Markov chain **initial values** (i.e. $\boldsymbol{\theta}^0$) were required to be specified for all parameters. In theory, the choice of initial values has no effect on the later sampled values from the chain, assuming that the values chosen are compatible with the parameter (e.g. initial values for a variance should be positive). However, in practice convergence of the chain to the equilibrium distribution can be improved by the choice of values (Spiegelhalter *et al*, 2004:106). In this thesis, the initial values were specified so as to lie within the range of plausible values for the parameters.

This thesis was not intended as an investigation into the influence of the choice of **prior distribution** on any inferences made. In general, **non-informative** (or **reference**) prior distributions were used by specifying a vague probability distribution; often $Normal(0,1000^2)$. Where data were sparse, external knowledge was used to create more **informative** priors, but no attempt was made formally elicit prior beliefs. Of course, it would be possible to repeat the modelling illustrated in this thesis using more informative priors.

It is important to determine that the Markov chain has **converged** to the equilibrium distribution and that it is sampling from the whole of this distribution. Problems may arise because the Gibbs sampler produces only a sample from the equilibrium distribution, rather

than the distribution itself and also, since a Markov chain is used, the sampled values will be correlated. There has been much discussion whether the number of chains generated can help to overcome, or at least identify, these problems. Running a number of chains, starting from different initial values, and checking whether they all arrive at the same sampling distribution can help to identify a lack of convergence in the chains (Gelman and Rubin, 1992; Gelman, 1995; Brooks and Gelman, 1998; Spiegelhalter *et al*, 2004). An alternative argument has been made that if a single chain is run for long enough then, no matter how correlated the chain, a suitable sample will be obtained (Geyer, 1992).

No methods exist to show for certain that a chain has sampled from the whole of the posterior sample space, though there are techniques that can offer evidence that such a sample has not been obtained. Formal methods exist to invesigate the convergence of a chain, for example Best *et al* (1996). The approach taken in this thesis was to run multiple chains, from dispersed starting values, as an initial inspection of models. Convergence was then checked for by both visual inspection of the trace plots and by the calculation of the Brooks-Gelman-Rubin statistic R (Brooks and Rubin, 1998). The Brooks-Gelman-Rubin statistic is the ratio of the width of the central 80% interval of the pooled chains to the average width of the 80% intervals within each chain. This statistic is available in WinBUGS, where it is calculated in bins of length 50 (Spiegelhalter *et al*, 1999b). Good convergence properties are shown both by the convergence of R to the value 1 and by the convergence of the width of the intervals to stable, and equal, values (Brooks and Rubin, 1998). Once a model was found to possess good convergence properties a single chain was used (usually with 10,000 sampled iterations) in other runs of the model for simplicity. In addition, the **autocorrelation** within the chain was visually examined for evidence of slow mixing of the chain and, therefore, the possibility that the chain did not cover the whole sample space.

For any chain, the iteration values generated before the equilibrium distribution has been reached need to be discarded (**burn-in**). In this thesis, a 1,000 iteration burn-in was used, which could have been increased if it was felt that the equilibrium distribution had not been reached after that number of iterations.

Advantages of Bayesian statistical methods

The two main advantages in the use of Bayesian methods for statistical analysis are often given as the ability to formally include prior knowledge, or beliefs, in the modelling and the ability to make probability statements about estimated parameters (Berry and Stangl, 1996:8). It has also been argued that a Bayesian approach can be more flexible than frequentist

methods (Spiegelhalter *et al*, 2004:3). Of course, these points are not universally accepted as advantages, particularly the first two. Indeed, a glance through any medical journal will show that the vast majority of the statistical analyses in published papers use classical statistical techniques. In using either approach in this thesis, emphasis is placed on the estimation of effect sizes and confidence (credible) intervals rather than solely on estimating statistical significance (p-values) because, after all, as Tukey (1969:86) said:

> *"The physical scientists have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "When you pull it, it gets longer!" Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared."*

3.3 Statistical Methods

A range of statistical methods has been suggested to compare binary outcomes. These range from basic 'rough-and-ready' methods to sophisticated statistical models. A simple method based on the ranking of the units, and two graphical approached to summarizing the data, are outlined in this section and discussed.

3.3.1 Rank

Perhaps the simplest method of comparing providers is to list them in order of the value of the chosen outcome (Rabilloud *et al,* 2001; Jenkins and Gauvreau, 2002). Such an analysis for the TNS data is shown in Table 3.1 using the observed mortality rates (Table 2.5). However, the use of such reporting is problematic as the simple ranking gives no indication of the uncertainty surrounding a provider's position (Spiegelhalter, 2003; Langford, 1997).

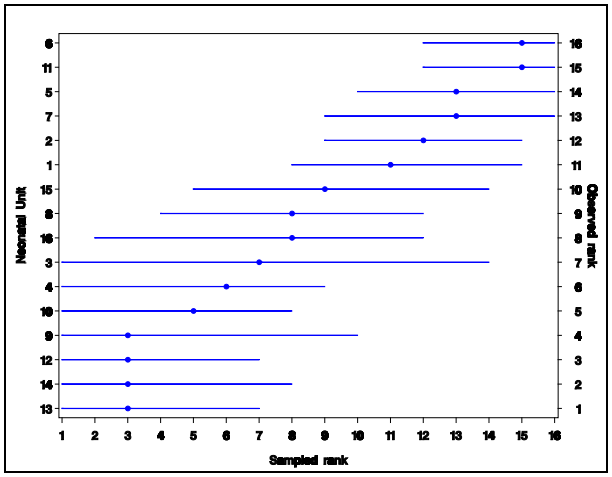
This problem can be solved by quantifying the uncertainty around each rank, either by using a bootstrap method or with a MCMC approach, by sampling the rank at each iteration and then reporting a confidence (or credible) interval from the sampled values (Marshall and Spiegelhalter, 1998b). These methods were applied to the TNS data. To obtain the bootstrapped estimates samples were drawn with replacement from the data for each NICU, with sample size equal to their number of observations. The mortality rates were then calculated for each NICU and each unit's rank calculated. This was repeated 1,000 times and then the 2.5^{th} and 97.5^{th} percentiles for each unit were estimated and reported as limits (Figure 3.1). The 50^{th} percentiles are also shown for information. The SAS/MACRO language was used and the macro ***boot_rank*** is shown in Appendix D.1. As the sampling distribution of

rank is unlikely to be symmetric, since it is bounded at both limits, it is recognised that this **percentile bootstrap method** may be an unreliable method to estimate a confidence interval (Chernick, 1999:53-54): more appropriate bootstrap methods are discussed in §5.6.3. However, it was felt that this, admittedly rather crude, method is sufficient to illustrate the problem with using ranks as a summary measure.

Table 3.1 NICUs ranked by rate of mortality

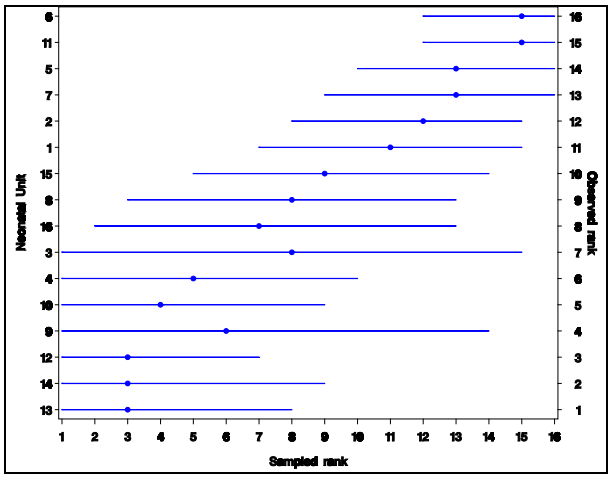
Rank	Unit	Total Infants	No. Died	Percentage died
1	**13**	136	3	2.2
2	**14**	90	2	2.2
3	**12**	196	5	2.6
4	**9**	35	1	2.9
5	**10**	146	5	3.4
6	**4**	142	6	4.2
7	**3**	38	2	5.3
8	**16**	100	6	6.0
9	**8**	124	8	6.4
10	**15**	124	10	8.1
11	**1**	212	21	9.9
12	**2**	283	30	10.6
13	**7**	243	29	11.9
14	**5**	333	41	12.3
15	**11**	445	62	13.9
16	**6**	378	54	14.3

Figure 3.1 Bootstrap 95% confidence intervals for rank



In the Bayesian approach, using MCMC sampling methods (in this case Gibbs sampling) a credible range for the ranks can be estimated in a similar matter to the bootstrap method outlined above, in this case sampling from binomial distributions with the parameters estimated using the observed values. After a 1,000 iteration 'burn-in', 10,000 iterations were performed using *Beta(2,18)* as the prior distribution for the sampled probability of death (to be discussed further in §5.3.2). These intervals are shown in Figure 3.2 and are, unsurprisingly given the similarity of methods, comparable to those in Figure 3.1.

Figure 3.2 Bayesian 95% credible intervals for rank



As can been seen in both Figure 3.1 and Figure 3.2 there is considerable uncertainty over the rank of any unit. Units 6 and 11 have the highest observed mortality rates but the 95% confidence (credible) interval for each of these units is from 12th to 16th place. Indeed, there is evidence that any one of seven of the units could truly have the lowest mortality rate. Such uncertainly is hidden when the ranks alone are reported.

In Chapter 5 a variety of outcome summary measures will be discussed and, in principle, they can all be used to rank the units. However, any risk adjustment methods used are likely to reduce the differences between the units, thus producing more uncertainty about the true rank of each unit.

3.3.2 Simple graphical approaches

Two graphical methods have been proposed to identify providers with extreme outcome rates: the **funnel plot** and the **spectrum plot**. Each of these is considered below.

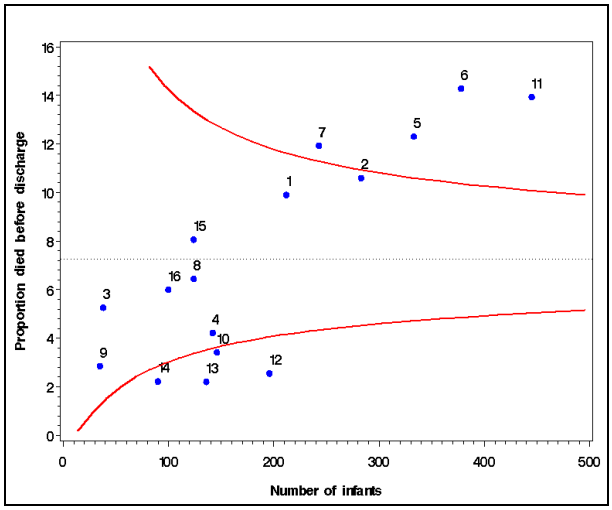
Funnel plot

One simple graphical method that has been proposed to illustrate observed rates (Stark *et al,* 2000) uses a graph known as a funnel plot, widely used in meta-analysis (Sutton *et al,* 2000:113). It is suggested that such a graph allows the examination of the rate of mortality by unit size, *"... providing a strong visual indication of 'divergent' performance or 'special cause' variation."* (Spiegelhalter, 2002)

An example, for the data used in this thesis, is shown in Figure 3.3 with a 95% exact binomial confidence limits shown (*2.1*) & (*2.2*), calculated assuming the mean Regional unit rate, i.e. $\frac{\sum_{j=1}^{16}\pi_j}{16} = 0.0729 = 7.29\%$.

Two points are immediately apparent from Figure 3.3. First, there is a trend towards higher in-unit mortality in the larger units. Second, some of the observed mortality rates fall outside the confidence interval shown. There are some units that appear to be performing better than the average and others that appear worse.

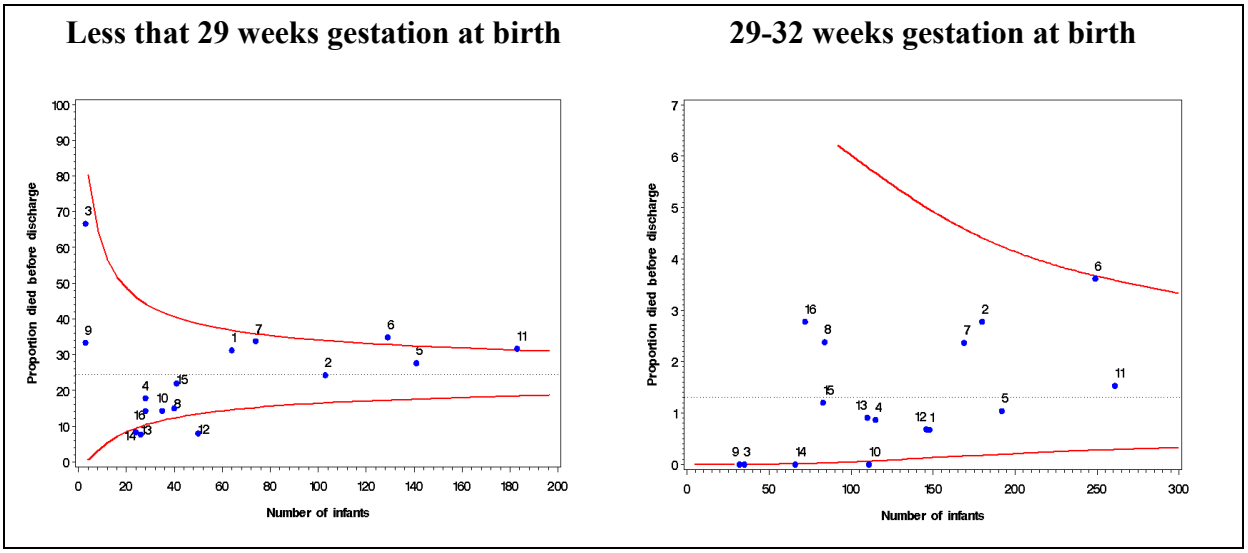
Figure 3.3 Mortality funnel plot for all admissions



While such plots only give limited information, they may be useful in the preliminary investigation of subgroups. For example, in §2.3 it was reported that a previous analysis using TNS data had shown different patterns of unit performance for those infants born at less that 29 weeks gestational age to those born between 29 and 32 weeks. Funnel plots for these two gestational age groups are shown in Figure 3.4.

Unit 6 has a high mortality rate for both gestational age groups, whereas Unit 11 only has a high rate for those infants with very low gestational ages.

Figure 3.4 Funnel plot for mortality by gestational age groups

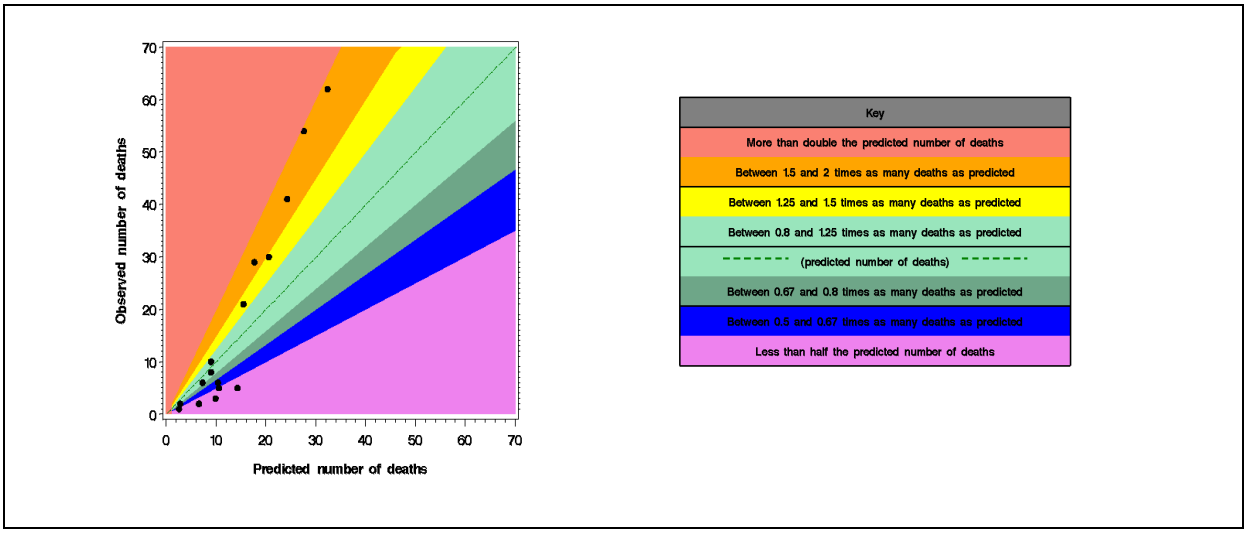


It has also been suggested that funnel plots be used to present the results of complex modelling of the type discussed in Chapter 5 (Simpson *et al*, 2003; Spiegelhalter, 2005). Such funnel plots become more difficult to produce once the outcomes are adjusted for case-mix since the size of the confidence intervals then becomes a function of not only the sample size but also of case-mix. One approach, used in a study of rates of severe intraventricular haemorrhage in neonatal units, is to interpolate between the estimated limits for each unit to form continuous (but not smooth) lines to represent the limits of the confidence intervals (Simpson *et al*, 2003). Such plots are useful in investigating any relationship between outcome and units size.

Spectrum plot

An alternative graphic is the spectrum plot, proposed by Heyward and Howman from the Medical Informatics and Clinical Governance Support Unit, University Hospital of Birmingham (Society of Cardiothoracic Surgeons of Great Britain and Ireland, 2002:207). In this example (Figure 3.5) the observed number of deaths is plotted against the number that would have occurred had the unit experienced the average Regional mortality rate of 7.29% (as derived above). The graph is divided into section according to the size of the ratio of the observed to predicted deaths. Although this graph is eye-catching, it does not give any indication of the uncertainty around the observations, unlike the funnel plot, and, therefore, is less useful.

Figure 3.5 Spectrum plot



3.4 Logistic Regression Models

While the methods outlined above may have some role in the preliminary inspection of the data, they do not provide estimates of effects and their associated uncertainty. In practice, statistical models are usually used for statistical estimation (Harrell, 2001:2). Such models naturally allow adjustment for case-mix differences between providers. Many of the models used in provider profiling can be seen to fall within the family of **Generalized Linear Models** (McCullagh and Nelder, 1989).

In order to express a response as a product of a linear combination, generalized linear models use a monotonic differentiable function (**link function**), $g(x)$, to map the predicted value to the interval $(-\infty, \infty)$. This function describes the relationship between the expected value and the **linear predictor** η.

Let $\mu_i = \mathrm{E}(\mathrm{Y} \mid x_i)$ $i = 1, \ldots, n.$

Then $g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ $\{\eta \in \Re\}$ *(3.6)*

The variance of the response depends on the mean through a **variance function** V;

$$\mathrm{var}(y_i) = \frac{\phi V(\mu_i)}{\omega_i} \qquad (3.7)$$

where: ϕ is the dispersion parameter;

$\boldsymbol{\omega}$ is a known weight for each observation: assumed for the rest of this thesis that $\omega_i = 1$, unless otherwise stated.

Generalized linear models for binary outcomes

Generalized linear models can be used for all outcomes through the appropriate choice of link function and variance function. However, in this thesis only the use of such models for binary outcome data will be pursued. Outcome variables that can take one of two possible values are common in provider profiling. Mortality is an often used outcome measure, e.g. in neonatal intensive care units (Parry *et al,* 1998) and following coronary bypass surgery (Peterson *et al,* 1998), but other examples of binary outcomes include complication after surgery (Silber *et al,* 1995), live birth rates from in vitro fertilisation (Marshall and Spiegelhalter, 1998b) and hospital readmission (Fisher *et al,* 1994; Daley *et al,* 1997).

Possible link functions for binary data include (where π is the probability of an event):

- the **logit** function: $g_L(\pi) = \log_e\left(\frac{\pi}{1-\pi}\right)$
- the **probit** function: $g_P(\pi) = \Phi^{-1}(\pi)$

  (where $\Phi(.)$ is the cumulative Normal distribution function)
- the **log-log** function: $g_{LL}(\pi) = -\log_e[-\log_e(\pi)]$
- the **complementary log-log** function: $g_{CLL}(\pi) = \log_e[-\log_e(1-\pi)]$

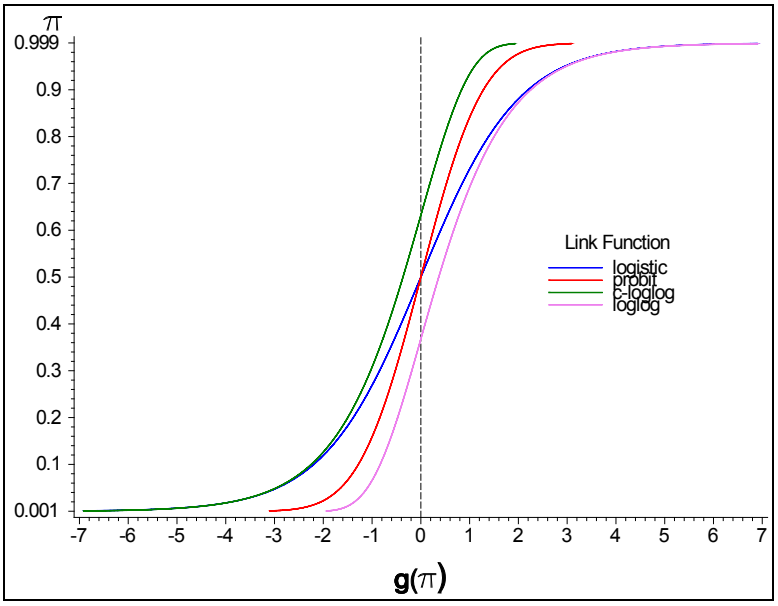
Each of these functions has the desired property of mapping the interval [0, 1] to the whole real line. However, there are both differences and similarities between these functions (Figure 3.6). For $0.1 \leq \pi \leq 0.9$ the logistic and the probit function are almost linearly related. When $\pi \leq 0.1$ the complementary log-log function and the logistic function are approximately equal, since they both approach $\log(\pi)$, and a similar argument holds for the log-log and logistic functions when $\pi \geq 0.9$. The logistic function and the probit function are both symmetrical in that:

$$g_L(\pi) = -g_L(1-\pi)$$

and

$$g_P(\pi) = -g_P(1-\pi).$$

Figure 3.6 Four possible link functions for binary data

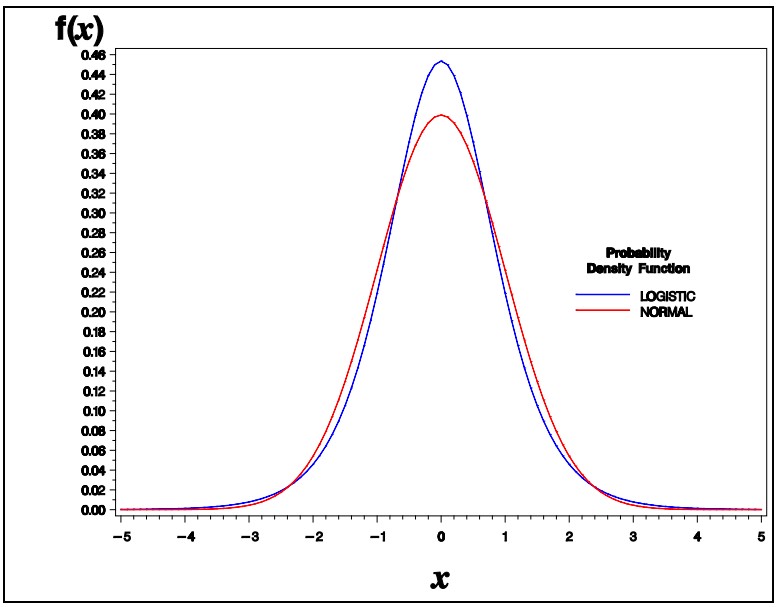


The log-log and the complimentary log-log functions are related:

$$g_{LL}(\pi) = -g_{CLL}(1-\pi)$$

In general, the logistic and probit link functions are likely to give very similar results: this is particularly true away from the tails. This is not surprising given the similarity between the two distributions (Figure 3.7).

Figure 3.7 Probability Density Functions of Logistic and Normal Distributions: Mean=0 and Variance =1



In practice, the logistic link is the most commonly used. The main reason for this choice is that the parameter estimates from a logistic model are easier to interpret: **log odds ratios**. However, there are additional reasons why, in general, the logistic link may be preferred. The logit transformation reflects an underlying qualitative variable (Binomial distribution), whereas the probit link reflects an underlying quantitative variable (cumulative Normal distribution) and it is also felt that logistic regression is more suited for the analysis of retrospectively collected data (McCullagh and Nelder, 1989:111).

This is not to say that the logistic link will be the most appropriate in all circumstances. However, the logistic link will be the first choice of link function for the rest of this thesis because of the availability of appropriate software (in particular SAS `PROC LOGISTIC`) and the greater knowledge of the properties of logistic regression models compared to models with alternative link functions.

The logistic regression model is given by:

$$\log_e\left(\frac{\Pr(Y=1\mid\mathbf{x})}{1-\Pr(Y=1\mid\mathbf{x})}\right)=\mathbf{x}'\boldsymbol{\beta} \qquad (3.8)$$

where: $\mathbf{x} = (1, x_1,\ldots, x_p)'$; the vector of explanatory variables;

$$Y=\begin{cases}1 & \text{if the outcome is positive}\\ 0 & \text{if the outcome is negative}\end{cases}$$

$\boldsymbol{\beta} = (\beta_0, \beta_1,\ldots, \beta_p)'$: the vector of unknown parameters.

The unknown parameters can be estimated by maximum likelihood.

Given: $P(Y=1\mid x)=\pi(x)$

then: $P(Y=0\mid x)=1-\pi(x)$.

Hence, the likelihood function is:

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}. \qquad (3.9)$$

and the log-likelihood is:

$$l(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \left(y_i \ln[\pi_i] + [1 - y_i] \ln[1 - \pi_i] \right) \qquad (3.10)$$

For models with p parameter estimates (p-1 covariates plus the intercept) this can be written as:

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\sum_{i=1}^{n_j} y_i x_{ij} \right) \beta_j - \sum_{i=1}^{n} \ln \left(1 + \exp \left[\sum_{j=1}^{p} \beta_j x_{ij} \right] \right) \qquad (3.11)$$

Partial differentiation with respect of each of the model parameters p gives the likelihood equations:

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n_k} y_i x_{ik} - \sum_{i=1}^{n_k} \hat{\pi}_i x_{ik} \qquad k = 1, 2, \ldots, p$$

Therefore, in matrix notation, the solutions are found by:

$$\mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\pi}}) = 0 \qquad (3.12)$$

The variances of the parameter estimates are given by the inverse of the information matrix $\mathbf{I}(\boldsymbol{\beta})$ where:

$$\mathbf{I}(\boldsymbol{\beta}) = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_m} = \mathbf{X}'\mathbf{V}\mathbf{X} \qquad k, m = 1, 2, \ldots, p \qquad (3.13)$$

where: $\mathbf{V} = diag\ \hat{\pi}_i(1 - \hat{\pi}_i)$

Usually this is estimated at the maximum likelihood estimates of the parameters ($\hat{\boldsymbol{\beta}}$). Hence, the estimated standard deviation of parameter β_k is given by the square-root of the k^{th} diagonal element of $\left[\mathbf{I}(\hat{\boldsymbol{\beta}})\right]^{-1}$.

Unlike linear regression, the maximum likelihood estimates for a logistic regression model cannot be written explicitly, except for special cases, such as when all the covariates are dummy variables (Harrell, 2001:228), and iterative methods are used to find solutions to the likelihood equations. The default method in SAS `PROC LOGISTIC` (the routine used in this thesis) is the Fisher-scoring algorithm, which is equivalent to estimation by iterative, reweighted least squares (IRLS). The alternative widely used approach is to use the Newton-Raphson method. In general, these two methods give the same parameter estimates but different standard errors. However, in the case of binary logistic regression the estimated

standard errors are identical, so the two methods give equivalent results (SAS Institute Inc., 1999:1904).

At the $(m+1)^{th}$ iteration the Fisher-scoring method estimates the expression (Agresti, 1990:449):

$$\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m + (\mathbf{I}_m)^{-1}\mathbf{q}_m \qquad (3.14)$$

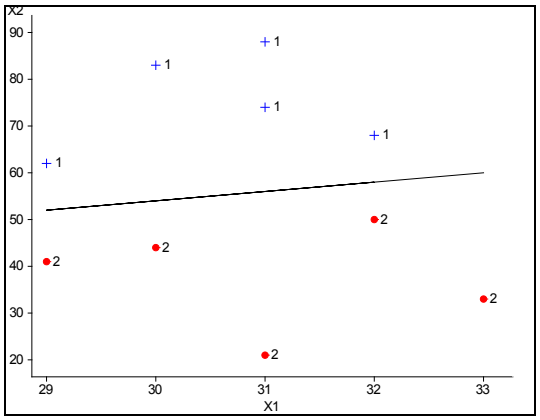
where: $\mathbf{I}_m$ is the information matrix evaluated at $\boldsymbol{\beta}_m$

$\mathbf{q}_m$ is the matrix $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \hat{\boldsymbol{\pi}}_m)$, where $\hat{\boldsymbol{\pi}}_m$ are estimated using $\boldsymbol{\beta}_m$

This iterative process continues until convergence is reached, that is, $|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m| < c$, where c is considered sufficiently small (the default in `PROC LOGISTIC` is $c = 1\times10^{-8}$). The estimates $\boldsymbol{\beta}_{m+1}$ are then the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$.

Although logistic regression models are now widely used, and standard options within most general statistical packages, there are potential problems. One of these is **separation**. If there is a covariate, or collection of covariates, that completely separate the outcome groups then there is said to exist **complete separation**. It can be shown that in this case the likelihood is monotonic and, therefore, the maximum likelihood estimates do not exist (Albert and Anderson, 1984; Bryson and Johnson, 1981; So, 1993).

Figure 3.8 Complete Separation



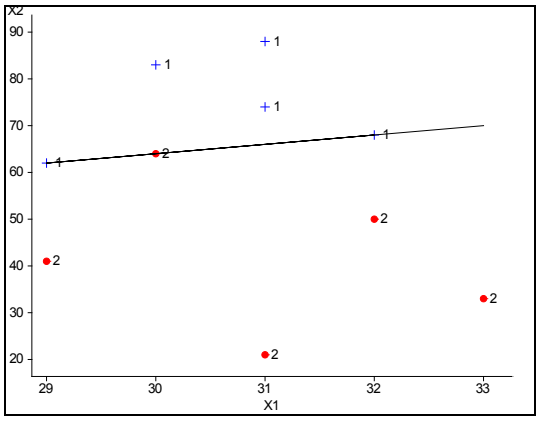
(adapted from So (1993))

Figure 3.8 shows the two outcome groups (1 & 2) completely separated, for example, by the line $x_2 = 2x_1 - 6$.

If there is not complete separation but the overlap is confined to a small number of tied values then **quasi-complete separation** exists. In this case there are solutions to the maximum

likelihood equations but typically the estimated standard errors of the coefficient estimates are very large.

Figure 3.9 Quasi-complete Separation

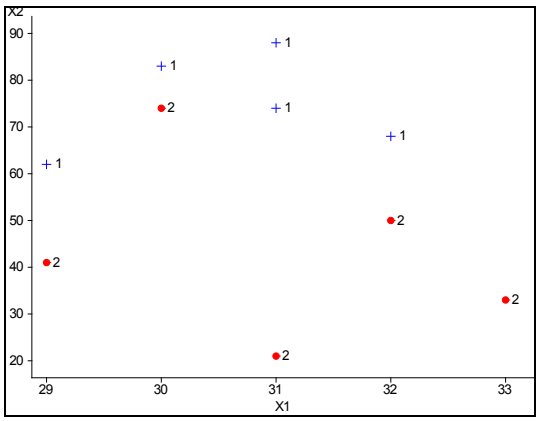


(adapted from So (1993))

In this case Figure 3.9 shows the two outcome groups cannot be completely separated.

In the absence of complete and quasi-complete separation there is an **overlap** between the two groups. An example is shown in Figure 3.10: it can be seen that the two groups cannot be separated by any straight line.

Figure 3.10 Overlap



(adapted from So (1993))

Complete and quasi-complete separation are more likely to be a problem with small samples, a small proportion of subjects with a particular outcome, and a large number of variables included in the model (Hosmer and Lemeshow, 2000:130).

SAS `PROC LOGISTIC` contains an empirical approach to try to identify complete and quasi-complete separation by reporting if the estimation procedure has not converged by the eighth iteration (SAS Institute Inc., 1999:1945).

3.4.1 Application to TNS data

Logistic regression model

The estimated log odds of mortality for each unit can be estimated using a logistic regression model. Using the overall model:

$$\log_e(\psi) = g = \beta_1 I_1 + \beta_2 I_2 + ... + \beta_N I_N \qquad (3.15)$$

where: N is the number of units

I_j is an indicator variable: $I_j = \begin{cases} 1 \text{ where infant allocated to Unit j} \\ 0 \text{ where infant not allocated to Unit j} \end{cases}$

β_j is the log odds for mortality for Unit j: i.e. $g_j = \beta_j$

This model is straightforward to estimate in SAS using `PROC LOGISTIC` by specifying that an intercept is not included in the model: `NOINT` (SAS Institute Inc., 1999:1920).

Confidence intervals for $\hat{g}$

The estimated confidence intervals for $\hat{g}$ are generally calculated using one of two approaches: Wald and likelihood ratio-based intervals (SAS Institute Inc., 1999:1950-1952).

The Wald confidence intervals are based on the assumption of asymptotic normality of the parameter estimates and the 100(1-α)% confidence interval for parameter β_j is given by:

$$\left(\hat{\beta}_j - z_{(1-\alpha/2)}\hat{\sigma}_j \ \textbf{\textit{to}}\ \hat{\beta}_j + z_{(1-\alpha/2)}\hat{\sigma}_j\right)$$

where: $\hat{\beta}_j$ is the maximum likelihood estimate of β_j

$\hat{\sigma}_j$ is the estimated standard error of $\hat{\beta}_j$ (see *(3.16)* and discussion)

The likelihood ratio-based intervals (also known as profile likelihood intervals) are based on the log-likelihood function. If, for example, a confidence interval for β_j is to be estimated, the profile likelihood function for $\beta_j = \theta$ is:

$$l_j^*(\theta) = \max_{\boldsymbol{\beta} \in \mathbf{B}_j(\theta)} l(\boldsymbol{\beta})$$

where: $\mathbf{B}_j(\theta)$ is the set of all $\boldsymbol{\beta}$ with the j^{th} element fixed at θ.

Let the log-likelihood evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ be denoted by $l_{\max}$, i.e. $l_{\max} = l(\hat{\boldsymbol{\beta}})$, then $2(l_{\max} - l_j^*(\beta_j))$ has a limiting chi-square distribution with one degree of freedom if β_j is the true parameter value. Therefore, limits to a 100(1-α)% confidence interval for parameter β_j is given by:

$$\{\theta : l_j^*(\theta) \geq l_0\}$$

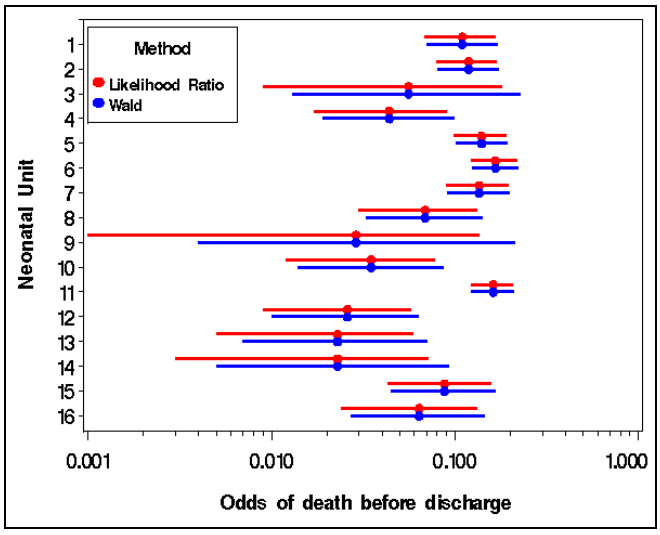
where: $$l_0 = l_{\max} - \frac{1}{2}\chi^2_{1-\alpha,1}$$

$\chi^2_{1-\alpha,1}$ is the 100(1-α) percentile of the chi-square distribution with one degree of freedom

The solutions are usually found by using an iterative algorithm (SAS Institute Inc., 1999:1950-1951).

Although confidence intervals based on the likelihood ratio are felt to be more accurate (Hosmer and Lemeshow, 2000:16; SAS Institute Inc., 1999:1950), they are computationally more complicated to obtain. Likelihood ratio-based intervals are not available from `CONTRAST` statements in `PROC LOGISTIC`, which is used later. To allow comparisons between different models Wald confidence intervals are the principal method used in this thesis: although Figure 3.11 displays both set of 95% confidence intervals for comparison.

Figure 3.11 Wald and likelihood ratio-base confidence intervals



For the large units the estimated confidence intervals were very similar for the two methods (Figure 3.11). However, when the units are small the forced symmetry of the Wald intervals

on the logarithmic scale was very apparent and this characteristic led to higher limits. Such intervals should, therefore, be interpreted with caution.

Estimated odds of death by unit

The results from the model specified by (*3.8*) are shown in Table 3.2 together with estimated Wald 95% confidence intervals

Table 3.2 Estimated odds of death before discharge for TNS data

Unit	Observed probability of death π_j	Odds of death $\hat{\omega}_j = \left(\frac{\pi_j}{1-\pi_j}\right)$	(95% Wald Confidence interval)
1	0.099	0.110	(0.070 to 0.173)
2	0.106	0.119	(0.081 to 0.174)
3	0.053	0.056	(0.013 to 0.231)
4	0.042	0.044	(0.019 to 0.100)
5	0.123	0.140	(0.101 to 0.195)
6	0.143	0.167	(0.124 to 0.223)
7	0.119	0.136	(0.091 to 0.200)
8	0.064	0.069	(0.033 to 0.142)
9	0.029	0.029	(0.004 to 0.215)
10	0.034	0.035	(0.014 to 0.087)
11	0.139	0.162	(0.123 to 0.212)
12	0.026	0.026	(0.010 to 0.064)
13	0.022	0.023	(0.007 to 0.071)
14	0.022	0.023	(0.005 to 0.093)
15	0.081	0.088	(0.045 to 0.168)
16	0.060	0.064	(0.027 to 0.146)

The maximum likelihood estimates for the odds were the same values as the observed. This can be seen more easily if the estimated probabilities of death are reported ($\hat{\pi}$), rather than the estimated odds ($\hat{\omega}$). These probabilities and Wald 95% confidence intervals are shown in Table 3.3. The estimated probabilities were equal to the observed values (Table 2.5) but the limits for the confidence intervals differ from the exact values calculated previously. The limits of the Wald estimated intervals took higher values than the exact intervals, although these differences were small.

Table 3.3 Estimated probability of death before discharge

Unit	No. infants	No. died	$\hat{\pi}_j$	(95% CI)
1	212	21	0.099	(0.065 to 0.147)
2	283	30	0.106	(0.075 to 0.147)
3	38	2	0.053	(0.013 to 0.187)
4	142	6	0.042	(0.018 to 0.090)
5	333	41	0.123	(0.091 to 0.162)
6	378	54	0.143	(0.111 to 0.181)
7	243	29	0.119	(0.084 to 0.166)
8	124	8	0.065	(0.032 to 0.123)
9	35	1	0.029	(0.004 to 0.176)
10	146	5	0.034	(0.014 to 0.079)
11	445	62	0.139	(0.110 to 0.174)
12	196	5	0.026	(0.010 to 0.059)
13	136	3	0.022	(0.007 to 0.066)
14	90	2	0.022	(0.005 to 0.084)
15	124	10	0.081	(0.043 to 0.143)
16	100	6	0.060	(0.027 to 0.127)

3.4.2 Bayesian estimates

Although exact confidence intervals for the probability of death before discharge can be calculated (see Table 2.5), one important advantage of using Bayesian methods is the ability to formally include prior knowledge into model. This is illustrated in this subsection.

The choice of a prior distribution is important, especially with small numbers, as any prior will have an influence on the posterior distribution since:

$$P(\mathbf{\Pi} \mid data) \propto L(\mathbf{\Pi} \mid data)P(\mathbf{\Pi}) \qquad \text{from } (3.2)$$

So while a specified prior distribution may be 'vague', it will never be completely 'non-informative' (Fisher, 1996). In this thesis the situation is complicated by the fact that the data have already been shown (Table 2.5). However, the discussion in this Section is based on knowledge from before this information was known.

It was assumed that:

$$d_{ij} \sim bernoulli(\pi_j)$$

where: d_{ij} is an indicator for death before discharge for the i^{th} infant at the j^{th} unit;

and: π_j is the probability of mortality before discharge at the j^{th} unit.

Hence, a prior distribution was sought for π_j.

Since $0 \leq \pi \leq 1$ two distributions were considered:

- $\pi \sim Uniform(\alpha, \beta)$

$$f(\pi \mid \alpha, \beta) = \frac{1}{(\beta - \alpha)} \qquad 0 < \alpha < \pi < \beta < 1$$

- $\pi \sim Beta(\gamma, \delta)$

$$f(\pi \mid \gamma, \delta) = \pi^{\gamma-1}(1-\pi)^{\delta-1} \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \qquad 0 < \pi < 1$$

The choice of *Uniform*(0,1), or equivalently *Beta*(1,1), as prior distribution for π quantifies the *a priori* belief that any value of π between 0 and 1 was equally likely. However, this was not the case since, at the very least, earlier data were available from the Trent Neonatal Study: between 1997 and 1999 the overall mortality in an equivalent population of admitted infants was 11.7% (The Trent Infant Mortality and Morbidity Studies, 2000). This knowledge could be used to inform the choice of the prior distribution. One option was to assume that a mortality rate of over 50% was impossible, but that values below this are equally likely, achieved by using *Uniform*(0,0.5) as the prior. An alternative approach was to acknowledge that not all values of π within a given range were equally likely. The use of the Beta distribution allowed this: the mean and variance of the beta distribution are:

$$E(\pi) = \frac{\gamma}{\gamma + \delta} \qquad (3.16)$$

$$Var(\pi) = \frac{\gamma\delta}{(\gamma+\delta)^2(\gamma+\delta+1)} \qquad (3.17)$$

It was felt, *a priori*, that mortality had fallen between the two time periods to around 10%. However, this is the overall mortality and it is known that rates for individual units vary from this. The values of γ and δ were varied to obtain suitable distributions.

The first Beta distribution was selected to have a mean value of 0.1. It was also felt that it was unlikely, although not impossible, that a unit would have a rate of over 25%. The distribution chosen based on these assumptions was *Beta*(2,18), giving:

$$E(\pi \mid \gamma = 2, \delta = 18) = 0.10$$

$$Var(\pi \mid \gamma = 2, \delta = 18) = 0.0043$$

$$P(\pi < 0.25 \mid \gamma = 2, \delta = 18) = 0.969$$

A second Beta distribution was chosen, *Beta*(4,28), so that the mode (rather than the mean) took the value 0.1 (mode = $\frac{\gamma-1}{\gamma+\delta-2}$), but that $P(\pi<0.25)$ was similar to the previous Beta distribution:

$$E(\pi \mid \gamma = 4, \delta = 28) = 0.125$$

$$Var(\pi \mid \gamma = 4, \delta = 28) = 0.0033$$

$$P(\pi < 0.25 \mid \gamma = 4, \delta = 28) = 0.969$$

However, the smallest observed values for π were 0.022 (Units 13 and 14) and the distribution *Beta*(4,28) gave inappropriately low probabilities to very small values of π: $P(\pi<0.02) = 0.0033$.

Therefore, a further Beta distribution was proposed that gave the maximum value for $P(\pi<0.02)$ while still having a modal value of 0.1: *Beta*(1.25, 3.25): further details are given in Appendix C.1. The distribution *Beta*(1.25, 3.25) has the following properties:

$$E(\pi \mid \gamma = 1.25, \delta = 3.25) = 0.278$$

$$Var(\pi \mid \gamma = 1.25, \delta = 3.25) = 0.036$$

$$P(\pi < 0.25 \mid \gamma = 2, \delta = 10) = 0.51$$

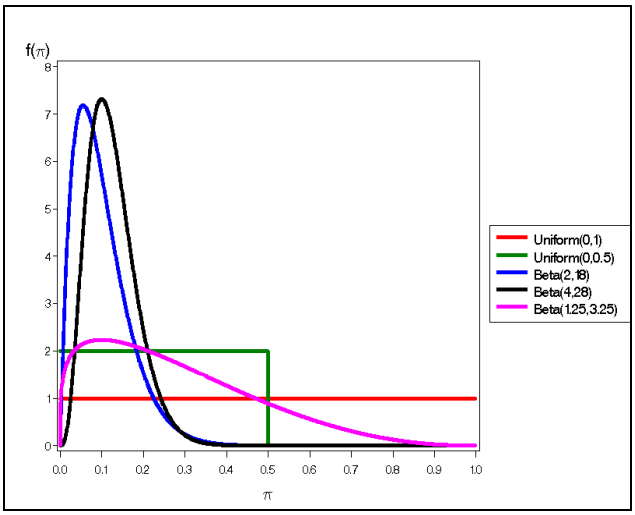
$$P(\pi < 0.02 \mid \gamma = 2, \delta = 10) = 0.030$$

The cumulative distribution functions of these distributions are shown in Figure 3.12.

To investigate the influence of the choice of prior distribution on the estimates, each of the distributions described above was used in a simple model to estimate the probability of death. Three units were investigated, one small, one medium and one large: Units 3, 1 and 5 respectively. Although it was recognised that this was not a rigorous investigation into the

influence of the prior distribution, it did offer some indication of the likely impact of a particular choice. Such a model can be expressed simply in WinBUGS (Appendix E.1).

Figure 3.12 Probability distribution functions of prior distributions



A 1,000 iterations 'burn-in' was inspected over five chains, using the methods described in §3.2.2, for reassurance that the markov chain had converged to the correct sampling distribution (examples of the diagnostic plots are shown in Appendix E.1). There was no evidence that the chains were not sampling from the full target distribution. The parameters of interest were then estimated from a further 10,000 iterations. The results are shown in Table 3.4.

Inspection of Table 3.4 illustrates two points. First, the choice of prior distribution was less influential for the larger unit (Unit 5) than for the other two. Second, different specifications of the Beta distribution produced different estimates for Units 1 & 3. Neither of these characteristics was unexpected: indeed it can be argued that they are both desirable. Where there is little information about the current performance of a unit (likelihood) then it seems reasonable that more weight will be given to prior knowledge (prior distribution). It has been suggested that trying to measure the performance of a health care provider over such a relatively short period of time "*... is rather like watching a football match for 10 minutes (randomly selected) and deciding that, if one team scores in this period, then it will win the match*" (Langford, 1997). To continue this analogy, surely we would want to take into account the score at the point we started to watch the match and, stretching it even further, we would want to allow for our prior knowledge of the teams.

The second point shows that when such knowledge is formally included in the model, care must be taken when specifying the prior distribution. The three Beta distributions illustrated

above are all compatible in some way with the prior information, but each produces different estimates for the smaller units. However, one of the advantages of Bayesian methods is that prior knowledge, or beliefs, can inform the posterior probabilities.

Table 3.4 Bayesian prior probabilities

Prior	Unit	No. Infants	No. Died	$\hat{\pi}_j$	WinBUGS Estimates		
					Mean	Median	(95% credible interval)
Uniform(0,1)	1	212	21	0.0991	0.1028	0.1011	(0.0664 to 0.1482)
	3	38	2	0.0526	0.0743	0.0672	(0.0163 to 0.1733)
	5	333	41	0.1231	0.1252	0.1245	(0.0925 to 0.1628)
Uniform(0,.5)	1	212	21	0.0991	0.1027	0.1018	(0.0660 to 0.1463)
	3	38	2	0.0526	0.0756	0.0680	(0.0162 to 0.1770)
	5	333	41	0.1231	0.1254	0.1246	(0.0919 to 0.1633)
Beta(2,18)	1	212	21	0.0991	0.0991	0.0979	(0.0645 to 0.1406)
	3	38	2	0.0526	0.0689	0.0636	(0.0190 to 0.1467)
	5	333	41	0.1231	0.1219	0.1211	(0.0897 to 0.1583)
Beta(4,28)	1	212	21	0.0991	0.1020	0.1008	(0.0681 to 0.1420)
	3	38	2	0.0526	0.0861	0.0823	(0.0324 to 0.1603)
	5	333	41	0.1231	0.1232	0.1223	(0.0914 to 0.1588)
Beta(1.25,3.25)	1	212	21	0.0991	0.1031	0.1019	(0.0666 to 0.1466)
	3	38	2	0.0526	0.0744	0.0675	(0.0163 to 0.1717)
	5	333	41	0.1231	0.1255	0.1247	(0.0921 to 0.1622)

It is possible to specify different prior distributions for the individual units. However, this will create difficulties in the context of provider profiling as choosing a prior distribution with its location more extreme that those of the other units can move the posterior distribution away from the others. While this is a possible approach to take, it would complicate the interpretation of any results: is an outlier the result of the chosen prior distribution? In this thesis, the same prior was used for all units as the prior belief was that all units were performing equally well, apart from sampling variation.

In Table 3.5 the mean sampled probabilities of death, with 95% credible intervals, estimated using *Beta*(2,18) as the prior distribution for all π_j. The point estimates and confidence intervals obtained were very similar to the frequentist results shown earlier (Table 3.3).

Table 3.5 Bayesian estimated probability of death before discharge for TNS data

Unit	Observed probability of death $\hat{\pi}_j$	Mean	(95% Credible interval)
1	0.099	0.099	(0.064 to 0.141)
2	0.106	0.106	(0.073 to 0.143)
3	0.053	0.069	(0.019 to 0.147)
4	0.042	0.049	(0.021 to 0.087)
5	0.123	0.122	(0.089 to 0.159)
6	0.143	0.141	(0.108 to 0.176)
7	0.119	0.118	(0.082 to 0.160)
8	0.064	0.070	(0.035 to 0.118)
9	0.029	0.055	(0.011 to 0.129)
10	0.034	0.042	(0.017 to 0.078)
11	0.139	0.138	(0.107 to 0.171)
12	0.026	0.032	(0.013 to 0.061)
13	0.022	0.032	(0.010 to 0.064)
14	0.022	0.037	(0.010 to 0.080)
15	0.081	0.083	(0.043 to 0.134)
16	0.060	0.066	(0.029 to 0.118)

A more intuitive approach may be to accept that there are different rates between the units and to specify a probability distribution for them. Such an approach is discussed briefly in §5.10.

3.5 Other Methods

The logistic regression model discussed in the previous Section offers a suitable approach to the analysis of the TNS data in this thesis, and the use of such models is developed further in Chapter 5. However, other statistical methods exist that may be appropriate for provider

profiling using binary outcomes but which are not suitable for the data to be analysed in this thesis. Such statistical methods are briefly discussed below.

Survival Analysis Techniques

One approach that may be considered is to use 'time to event' (survival) methods. However, such methods are unlikely to be appropriate in this case. First, there are no censored observations, except possible cases of infants discharged home when thought to be terminally ill. However, such cases are rare in neonatal medicine. Second, and more importantly, small increases in survival amongst neonates are not necessarily appropriate. It is felt that it is not always in the best interest of the infant, and the parents, to extend life at any cost. The Royal College of Paediatrics and Child Health recommend five situations where it is appropriate to consider withholding or withdrawing treatment in children (Royal College of Paediatrics and Child Health, 1997:7). Three of the situations are particularly relevant to neonates (the other two are "*The Brain Dead Child*" and "*The Permanent Vegetative State*"):

> *"* ***3. The 'No Chance' Situation.*** *The child has such severe disease that life sustaining treatment simply delays death without significant alleviation of suffering. Medical treatment in this situation may thus be deemed inappropriate.*
>
> ***4. The 'No Purpose' Situation.*** *Although the patient may be able to survive with treatment, the degree of physical or mental impairment will be so great that it is unreasonable to expect them to bear it. The child in this situation will never be capable of taking part in decisions regarding treatment or its withdrawal.*
>
> ***5. The 'Unbearable Situation.*** *The child and/or family feel that in the face of progressive and irreversible illness <u>further</u> treatment is more than can be borne. They wish to have a particular treatment withdrawn or to refuse further treatment irrespective of the medical opinion on its potential benefit. Oncology patients who are offered further aggressive treatment might be included in this category. "*

There are few data on the proportion of deaths that follow an informed and agreed decision to withdraw treatment. One study reported that 30% of deaths in one London neonatal unit between 1982 and 1985 followed the withdrawal of treatment (Whitelaw, 1986). A recent study from France found that 44% of deaths on neonatal units followed decisions to withhold or withdraw treatment (Larroque *et al,* 2004), and an earlier study reported that 14% of neonatal deaths at Yale-New Haven Hospital, Connecticut USA between 1970 and 1972 followed the withdrawal of treatment (Duff and Campbell, 1973). In the Netherlands, where active euthanasia is legal in some circumstances, it is not legal in the case of a severely disabled baby as such action must be requested by the patient. However, it has been suggested that about 100 babies die each year as a result of decisions made by doctors to

hasten their death (Sheldon, 2004). It has also been shown that clinicians' attitudes to active euthanasia differ widely, with some 20% of UK neonatologists surveyed stating that they believed that the current law should be changed to allow active euthanasia "more than now" (Cuttini *et al,* 2004). While this is not to say that such practices occur in British neonatal units, it is an indication that quality of life is an important consideration taken together with survival.

Although it is unknown whether these decisions ultimately affected whether these infants survived until discharge, it is clear that they influenced the length of survival. Using the length of survival as an outcome measure, that is using survival analysis methodology, implies that it is this that is important rather than the quality of care. This is clearly inappropriate. For other outcomes, there may be a case for using such statistical methods. The time that a particular interventions is used, for example mechanical ventilation, may lend itself to these techniques, although there may be problems with informative censoring as the sickest infants are more likely to die while on treatment and, therefore, be censored observations.

Ecological Regression

In cases where individual patient data are not available, summary data may be used in a linear regression model (Joyce *et al,* 2002). Such methods are used by The Dr Foster organisation in its published "Guides" to hospital and consultant outcomes (Dr Foster, 2004). The ratio of the observed number of deaths to the expected number of deaths (indirectly standardized to the whole population) is used as the outcome in a weighted linear regression model. Variables used as covariates in such models include *"aggregated discharge data such as the percentage of emergency cases, individual hospital data such as total number of beds, and community attributed data such as the percentage of patients with limiting longstanding illness"* (Jarman *et al,* 1999).

However, the use of aggregated data presents distinct problems. In particular, it is not certain that a correlation that exists within a group also exists at the individual level (Robinson, 1950). The assumption that such a relationship exists is often known as the **ecological fallacy** (Selvin, 1958). A further problem with such modelling is that the use of aggregated data means that there are often only a few data points (only the number of units being investigated) and, therefore, there is little statistical power to investigate relationships within the data (Lambert *et al*, 2002, Sterne *et al,* 2001).

Since individual level data are available in this thesis, and because of the problems outlined above, such modelling will not be investigated in this thesis.

Continuous Monitoring

Various methods have been advocated, and used, that involve continuous updating of outcome summaries. Some of these derive from methods used in industrial process control, for example Shewart's method (Braitman and Davidoff, 1996; Adab *et al,* 2002; Mohammed *et al*, 2001a), sequential probability ratio test (SPRT) (Spiegelhalter *et al,* 2003) and cumulative sum (CUSUM) (Poloniecki *et al,* 1998). Others are derived from clinical data, for example the variable life-adjusted (VLAD) plot (Lovegrove *et al,* 1997; Sherlaw-Johnson *et al,* 2000).

However, these methods will not be reviewed further in this thesis. The nature of the data collection methods used with the Trent Neonatal Survey make such methods difficult to implement. The data are collected from the units by research nurses, who may only visit small units every few months. In addition, and perhaps more importantly, the data are not entered onto the database until the end of the calendar year. Therefore, any use of continuous monitoring would be retrospective, this removing any advantage such methods may have over the cross-sectional approach taken here.

Perhaps these methods are most suited to monitoring the performance of a single unit, reflecting the process control role many of them originate from.

Conjoint Analysis

Conjoint analysis can be used to elicit preferences and may have a role in provider profiling. The essence of the method is that various scenarios are proposed which are then ranked or rated. These results can then be used in a linear regression model where changes in preference derived from a change in a characteristic can be estimated (Ryan and Farrar, 2000). However, such a technique is most likely to be of benefit in profiling where several characteristics of a provider have been estimated and their relative importance for classes of users is to be investigated.

Bradley-Terry Model

The Bradley-Terry method is used to model paired preference data (Bradley and Terry, 1952). Such preferences could then be used to compare institutions (Dittrich *et al,* 1998). Using the

TNS data, preference probabilities could be obtained for all $\binom{16}{2}$ pairs among the 16 units. For each unit a parameter π can be defined such that $\pi_k \geq 0$, k = 1,…, 16 and $\sum_{k=1}^{16} \pi_k = 1$ and so that the probability that unit *x* is preferred over unit *y* is $\frac{\pi_x}{\pi_x + \pi_y}$ for all *x* and *y*.

The maximum likelihood estimates for the unit parameters (i.e. $\hat{\pi}_1, \hat{\pi}_2 ..., \hat{\pi}_{16}$) can be shown to be the solutions to:

$$\hat{\pi}_j = \frac{a_j}{\sum_{\substack{k=1 \\ k \neq j}}^{16} \left(n_{jk} \Big/ \left[\hat{\pi}_j + \hat{\pi}_k \right] \right)}, \qquad \text{subject to } \sum_{j=1}^{16} \hat{\pi}_j = 1 \qquad \textit{(3.18)}$$

where a_{jk} is the number of times unit *j* is preferred over unit *k* in n_{jk} comparisons and $a_j = \sum_{\substack{k=1 \\ k \neq j}}^{16} a_{jk}$ (Kotz and Johnson, 1982:313).

While this approach of comparisons between units may be appealing, it is not clear how such models can be applied using TNS data. In the case of the data in this thesis no direct comparisons between the units have been made. It may be possible to imply such comparisons from the observed outcomes of similar infants, but such an approach is unlikely to offer any advantages over logistic regression modelling in this case. This approach will not, therefore, be considered further here.

3.6 Chapter Summary

The statistical methods in this chapter have been proposed to quantify the sampling variation in binary performance indicators. In §3.3 some simple methods that may be useful in preliminary investigations were illustrated. It was argued that these methods were insufficient for a robust investigation, but that logistic regression models were sufficiently flexible to be of most use in provider profiling. These models were introduced in §3.4 and illustrated using both Classical and Bayesian approaches. Other statistical methods proposed for provider profiling, but unsuitable for the data in this thesis, were briefly described in §3.5.

As discussed in §1.3.2, there is also uncertainty in the mortality rates due to differences in morbidity between the infants in different units. Any analysis that does not take such differences into account is likely to produce biased results. This is certain to be true with the TNS data in this thesis as the larger 'lead' units are likely to have sicker infants than the other units. It is this source of uncertainty that is discussed in the next Chapter.

Chapter 4: MORTALITY RISK ADJUSTMENT

4.1 Chapter Overview

Section 4.3 introduces risk scores. Scores specifically designed, or advocated, to quantify the mortality risk of neonates are described in §4.4, and studies that have compared these scores are discussed in §4.5. The approach to risk-adjustment taken in this thesis is described in §4.6, while §4.7 sets out the main conclusions of the Chapter.

4.2 Background

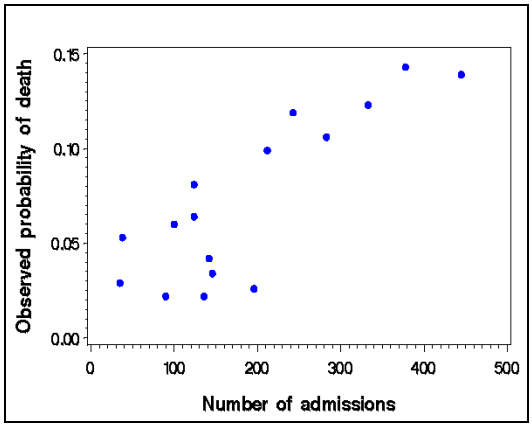
It was suggested in §1.3.2 that some of the variation in mortality rates between the units may be due to differences in the morbidity of the infants, rather than the care given. This potential for bias was described in 1864 (in response to a letter from William Farr):

> *"Any comparison which ignores the difference between the apple-cheeked farm-labourers at Stoke-Poges (probably for rheumatism and sore legs), and the wizened [sic], red-herring-like mechanics of Soho or Southwark, who come from a London Hospital, is fallacious."* (cited in Iezzoni, 1997)

This is particularly important with TNS data, as it was expected that there would be clinical differences between the infants in referral units and those in local units. Since the smaller units will transfer out many of their sickest infants to the referral units, these smaller units are likely to have populations with relatively better prognoses. This is indeed the case (Figure 4.1) with the larger units also having the highest observed mortality rates (from Table 2.5).

Differences in case mix between units have also been noted in adult intensive care and, moreover, it was observed that such differences were associated with hospital mortality (Rowan *et al,* 1993b).

Figure 4.1 Observed mortality by unit size



The process of formally allowing for differences in patients' characteristics has a number of names, for example case-mix, severity, sickness, intensity, complexity, comorbidity, health status (Iezzoni, 1994). The term generally used, and the one used in this thesis, is **risk-adjustment**. However, it is important to be clear what risk is being adjusted for. A 'one-size-fits-all' approach is unlikely to work and any adjustment needs to be adapted to the outcome of interest.

One simple method of risk-adjustment is to compare only units with similar types of patients, an approach suggested for comparing general practitioners (Baker *et al,* 2003). However, this is difficult to achieve, perhaps impossible, where there are a small number of units, as in TNS. One study sought to find groups of obstetric units with similar case-mixes using cluster analysis amongst 159 hospitals in Bavaria, Germany but was unable to find any such groups (Selbmann *et al,* 1982). Therefore, an alternative approach is required that tries to quantify the mortality risk in individual infants.

Various risk-scores have been proposed to quantify the morbidity of an individual and, when appropriate, these are useful tools. The scores discussed in this thesis are designed to investigate the binary outcome of death before discharge or alive at discharge. It is quite possible for scores to be used to try to estimate the probability of types of outcomes other than binary, such as the Cambridge Baby Check Score Card, which categorises infants under the age of six months into bands of illness severity (Morley *et al,* 1991). However, as mortality is the outcome used in this thesis, such scores are not discussed.

While the use of risk-adjustment is generally recognised as appropriate, there is an argument that the reporting of risk-adjusted indicators suggests an inappropriate level of accuracy. It is argued that the reporting of unadjusted rates, with a more strict definition of extreme

performance, makes it explicit that the indicators do not represent the 'truth' but are a guide for further investigation (Keogh *et al,* 2004). However, adjusted rates are a step towards the 'truth', and all provider profiling should be taken as screening for providers worthy of further investigation. This thesis, therefore, considered risk-adjusted outcomes.

4.3 Risk Scores

The use of a **risk score**, calculated using patient characteristics, to predict the probability of an event is the most commonly used approach to risk adjustment. The desirable properties of a neonatal score have been described as including: "*(1) ease of use; (2) applicability early in the course of hospitalisation; (3) ability to reproducibly predict mortality, specific morbidities, or cost for various categories of neonates; (4) usefulness for all groups of neonates to be described*" (Fleisher *et al,* 1997). It has also been said that *"the whole trick is to decide what variables to look at and then to know how to add"* (Dawes and Corrigan, 1974). Assuming that adding is not a major problem, the greatest difficulty arises in selecting the variables to be included in the model. Investigators will often require a ready made and validated risk adjustment method as they may lack the data, resources, time, funding or expertise required to develop their own score (Rosenthal and Harper, 1994). A previously validated score also has the advantage that the results of any analysis are more like to be accepted by others and more easily allow comparison across studies.

4.3.1 Methods for deriving a risk score

Risk scores are generally produced in one of two ways:

- **Medical models**: These are derived using clinical knowledge and observed behaviour. Many of the early scores produced in the 1970s and 1980s were medical models as there were few large datasets available to developers (Iezzoni, 1997);
- **Data models**: Collected data are used to produce these risk scores. The variables included, and the values for the coefficients, are chosen according to some statistical criterion. Most scores developed in recent years are derived using this approach, although medical knowledge may have, indeed should have, contributed to the choice of variables to be included in the model.

There are merits to both approaches. As implied above, **medical models** can be derived where there is an absence of previously collected data. Such models may well also attract more confidence from clinicians than statistically derived models. However, there is evidence that statistically derived **data models** have superior predictive ability (Dawes, 1980; Einhorn, 1986). However, it is useful if the model is clinically plausible, not for an absolute statistical reason as good prediction is sufficient, but rather so that providers are convinced that risk has been adequately allowed for (Iezzoni, 1994).

4.3.2 Use of risk scores

If a suitable pre-existing risk score can be identified, it can be used in different ways. When the coefficients (raw or occasionally rounded) from the logistic regression model are reported by a score's developers, these coefficients, forming the linear predictor derived from the developers' data, can then be applied to the sample of interest. When a logistic regression model has been used, the estimated probabilities of an event for each observation can then be obtained using the inverse logit transformation. For example, the authors of the Berlin Score (discussed further in §4.4.5) published the following linear predictor (Maier *et al,* 1997):

$$\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -3.7 - \left(0.84 \times BW_i\right) + \left(0.71 \times RDS_i\right) - \left(0.55 \times APGAR_i\right) + \left(0.53 \times VENT_i\right) + \left(0.37 \times BE_i\right)$$

where: BW = birth weight group;

RDS = grade of respiratory distress;

$APGAR$ = Apgar score at 5 minutes;

$VENT$ = administering of artificial ventilation;

BE = base excess at admission.

The probability of death π_i can be estimated from this function. In fact, this is indirect standardization (discussed in more detail in §5.4) with the sample used to estimate the logistic model as the reference population.

An alternative approach is for the coefficients, usually in this case rounded, to be used to form a score for each observation and this score then to be entered as a covariate in a new logistic regression model using the data of interest. The individual probabilities of an event are then estimated from this second model. As well as presenting the linear predictor from their

logistic regression model, the authors of the Berlin Score also published details of how to calculate a single score value from estimated values of the parameter estimates (details in Appendix B). In this case, the score can take a value from 0 to 40 with higher scores representing higher morbidity. This score can then be used as a covariate in a new logistic regression model. Mortality rates change over time and a score may become poorly calibrated over time. Using the score in a new model ensures that the observed and expected number of deaths are the same, thereby recalibrating the model so that it better represents the new data (Pollack *et al,* 2000; Ivanov *et al,* 1999). Such an approach also allows additional variables to be added to the model if this was important.

If it were felt that the published linear predictor was unsuitable (since, for example, surfactant had not been administered to infants in the original study) then another approach would be to estimate a new linear predictor using the variables identified in logistic regression model. This would also allow the introduction of additional variables.

Which approach is most appropriate in any particular circumstances will depend on the question to be answered and the data available. For this thesis, data are available to be used as a reference for each unit. The preferred approach would be to use an existing score as a covariate in a new model. This allows the use of a pre-existing and validated mortality prediction method but also allows recalibration to current mortality rates.

Variables to be included in a neonatal score

If the aim of a study is to make some inference on the quality of care provided, as is the case in this thesis, risk-adjustment must not include any variables that this care could influence. There is evidence that data collected a short time (up to 24 hours) after admission produce better discriminating models than data collected solely at birth (Pollack *et al,* 2000). However, the inclusion of such variables in the score can cause problems with an inappropriately treated infant receiving a 'worse' score than an adequately treated infant and, because of this, having a higher predicted mortality (Boyd and Grounds, 1994). Such a unit would then appear better than it should, since its infants will seem to have poorer prognoses as the result of their early treatment. It is, therefore, important to recognise when an infant first comes under the care of the unit. Neonatologists are likely to have been involved in an infant's care before it physically arrives on the unit. Such circumstances allow the neonatal team the opportunity, unintentionally or intentionally, to influence an infant's risk score, and hence its predicted probability of death. The uptake of the recommendation that units should have guidelines for when senior or consultant neonatal staff should attend preterm births

(CESDI, 2003) would mean neonatal care beginning before the infant arrives on the neonatal unit more often. The policy of the hospital can also influence the estimated mortality ratio, as infants stabilized before transfer will have better signs when it arrives on the unit. The answer may be to investigate obstetric and neonatal care as one, although internal politics in hospitals may make this impossible to achieve.

An example of how an infant's condition can be affected by the neonatal team before its arrival on the unit is given by its temperature on admission. Temperature on admission is a known risk factor for neonatal mortality (Parry *et al,* 2003b). It can be appropriately controlled with good clinical practice (Lyon and Stenson, 2004), just as poor practice can produce inadequate temperature control (Rennie and Roberton, 2002:87):

> *"*[Transfer from the labour ward] *is where the control and care of the sick neonate often starts to go awry. It is a grossly substandard level of care – but one that applies all too often – to transfer a sick baby with inadequate respiration from the labour ward to the NNU wrapped in a blanket and without supplementary oxygen. The baby arrives in the NNU cold, blue, limp and grunting (if you are lucky) or apnoeic and half dead (if you are unlucky)."*

Project 27/28, a study of infants born at 26 to 29 weeks gestational age, found that 61% of admissions to a NICU did not achieve the minimum required temperature of 36ºC and that it took a median time of 2 hours (range 1.5 to 4 hours) for the infants' temperature to be corrected (Jain and Fleming, 2004). Hence, the inclusion of admission temperature in a risk score means that the quality of the early care provided by the neonatal team influences the predicted probability of death. This is an undesirable quality in this type of profiling.

There is evidence that clinicians' assessment of mortality risk can improve the performance of a neonatal risk adjustment model (Stevens *et al,* 1994). However, while this may be important in clinical practice, for example decisions to transfer or to decide on treatment, such methods are unsuitable for provider profiling for the obvious reason that this cannot be standardized across all providers.

Missing data

The problem of missing data for a risk score is often overcome by assuming that the missing values would be within the 'normal' range; for example Paediatric Index of Mortality (PIM) (Shann *et al,* 1997), Score for Neonatal Acute Physiology (SNAP) (Richardson *et al,* 1993), Medicare Mortality Predictor System (MMPS) (Daley *et al,* 1988). It is argued that

pathological or metabolic measurements are not routinely taken if it is felt that they will show no abnormality. Therefore, substituting values considered appropriate allows these observations to be used. While such an assumption may not true in all cases, it may offer a simple method to include all observations and it has been shown that the substitution of modal values (often 'normal' values) performs well in prognostic models (Ambler *et al*, 2005). However, if inappropriate, the substitution of 'normal' values means that this method will underestimate the morbidity of an infant and, hence, make it appear that the unit is performing worse that it really is. It could be argued that such an approach provides an incentive for rigorous data collection. A more robust approach may be to estimate the missing values using imputation methods (Zhang, 2003). However, such methods are often not straightforward to apply and were not used in this thesis.

Appropriateness of the score to sub-groups

Any method applied to the data should not just ensure a good overall fit to the data but should adequately describe subgroups too. For example, the USA-derived APACHE II equation, when fitted to British adult intensive care patients, found under-prediction of mortality in patients 76 years of age or older (Rowan *et al,* 1993a). The authors concluded that such differences may be due to several reasons: systematic differences in medical definitions and diagnostic labelling between the two countries; true differences in the diagnostic mix between countries; systematic differences in the measurement of physiological variables; systematic differences in the effectiveness of treatment; the possibility that differences exist in the age specific health status between the countries.

All scores make the assumption that the risk factors for each potentially fatal condition are the same but this may not be the case (Iezzoni, 1994). The ideal system may use a core group of acute physiological variables to which are added a small subset of condition-specific clinical variables (Iezzoni *et al,* 1992). However, such an approach is difficult with the small datasets available in neonatal medicine. In certain circumstances it may be appropriate to investigate some sub-groups separately because potential confounders cannot be collected for all patients. For example, in a study comparing three month outcome after stroke in twelve centres across Europe, separate models were used to estimate outcome depending on whether the patient was in a coma at the initial examination since, if so, data could not be collected at that time on variables such as swallowing and limb movement (Wolfe *et al,* 1999). However, such difficulties are unlikely to arise with the variables collected by TNS.

Model validation

Before any risk score can be used, confirmation is required that it works 'satisfactorily' for data other than those used to derive it (Altman and Royston, 2000). This is usually called **model validation**. This is acheived either by **internal validation** methods using sub-sets of the data to develop and test the model, such as data-splitting, cross-validation and bootstrapping, or by **external validation**, applying the model to independent data (Harrell *et al*, 1996).

This thesis does not attempt an investigation into model validation methods. However, both internal (§6.5) and external (§6.8) validation methods are applied to models in Chapter 6.

4.4 Neonatal Mortality Risk Scores

A variety of risk adjustment scores have been derived, or advocated, for assessing the risk of neonatal mortality. Scores identified from the literature are:

- Clinical Risk Index for Infants (CRIB) and CRIB II
- Score for Neonatal Acute Physiology – Perinatal Extension (SNAP-PE) and (SNAP-PE II)
- Neonatal Therapeutic Intervention Scoring System (NTISS)
- National Institute of Child Health and Human Development (NICHHD)
- Berlin score
- Sinkin scores
- Neonatal Mortality Prognosis Index
- Apgar Score
- Transport Risk Index of Physiologic Stability (TRIPS)

Each of these scores will be briefly described below, with further details of the variables used in each scoring system given in Appendix B.

4.4.1 Clinical Risk Index for Infants (CRIB)

The Clinical Risk Index for Infants (CRIB) is a widely used mortality risk score for neonates. It has recently been updated (Parry *et al,* 2003b), under the title CRIB II, to reflect changes in survival since the score was first published in 1993 (The International Neonatal Network, 1993). The original CRIB was developed using 812 infants born either with a birth weight of 1500g or less or at less than 31 weeks gestational age. The data were admissions to four UK tertiary hospitals between 1988 and 1990.

The variables in the final model were chosen from a pre-selected set of obstetric and neonatal variables using a logistic regression model. The estimated model coefficients were then converted to an optimally chosen set of integers (Cole, 1993). The score was designed to be used as a covariate in a locally derived logistic model, perhaps with the inclusion of additional important predictors. However, in a later publication the authors gave the linear predictor from their original model:

$$\log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -4.070 + 0.445 \times CRIB_i$$

(Scottish Neonatal Consultants' Collaborative Study Group and the International Neonatal Network, 1995)

Data on admissions to four other teaching hospital neonatal units over the same time period (n = 488) were used to validate the predictive ability of the score: area under the ROC curve $AROC = 0.90$; se = 0.05 (§6.3.1). Calibration was assessed using the Hosmer-Lemeshow goodness of fit test (discussed in more detail in §6.3.2): $\chi^2_{df=10} = 16.84$; p = 0.078. This test suggests that there may be some weak evidence for the model not predicting mortality probabilities well. This together, with the fact there was no attempt to investigate the performance of the model in subgroups of admissions, means that although the discriminatory ability of the model is good we can be less sure of its calibration. The only component in the score allowing for gestational age is whether the infant was born at 24 weeks gestational age or less. As gestational age at birth is arguably the single most important predictor of survival (see §6.4.1), this would seem to be insufficient adjustment. However, this could be overcome by including gestational age at birth as a covariate in a model together with CRIB.

Although CRIB was originally derived as a mortality score it has also been used to try to predict other outcomes: e.g. major impairment at 18 months of age (Scottish Neonatal

Consultants' Collaborative Study Group and the International Neonatal Network, 1995), retinopathy of prematurity (Vyas *et al,* 2000).

In 2003 the revised version of CRIB was published: CRIB II (Parry *et al,* 2003b). CRIB was originally developed before the routine use of surfactant was introduced to neonatal units. In addition, during the ten years between the publication of the two versions of the score, a growing number of very preterm infants (less than 26 weeks gestational age at birth) have been admitted to neonatal units. The original score had become poorly calibrated for neonatal mortality. In addition, it was recognised that the inclusion of observed maximum appropriate inspired oxygen concentration meant that differences in oxygen monitoring or administration can influence the calculated score (Baumer *et al,* 1997).

CRIB II was developed using 1886 infants admitted to 35 randomly selected neonatal intensive care units in the UK. The final score comprised an item derived from previously published mortality rates by gestational age, birth weight and sex (Draper *et al,* 1999), together with temperature at admission and base excess. This was validated using data from 1065 infants in 19 other NICUs: $A_{ROC} = 0.92$; $\hat{C} = 4.30 \sim \chi^2_8$, p = 0.83 (see §6.3). The authors' estimated linear predictor was:

$$\log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -6.476 + 0.450 \times CRIB\ II_i$$

The application of CRIB and CRIB II to the TNS data is taken up in §6.9.4.

4.4.2 Score for Neonatal Acute Physiology (SNAP)

The Score for Neonatal Acute Physiology (SNAP) was first proposed in 1993 as a measure of infant morbidity (Richardson *et al,* 1993). It was developed using data from 1643 admissions to three NICUs in the USA. The final score comprised 34 items for which data are collected over the first 24 hours of life. This score was then modified to include measures of low birth weight to produce the SNAP Perinatal Extension (SNAP-PE) to quantify mortality risk (Richardson *et al,* 1993).

The SNAP and SNAP-PE were both updated and simplified in 2001 using data on 10,819 admissions to NICUs in Canada (Richardson *et al,* 2001). The scores were validated using data from Canada, California and New England and were considerably simpler than the original scores. The revised scores also reduced the data collection period from 24 hours to the first 12 hours of life and found to fit these data well.

4.4.3 National Therapeutic Intervention Scoring System (NTISS)

The NTISS was derived by modifying the Therapeutic Intervention Scoring System (Cullen *et al,* 1974; Keene and Cullen, 1983), a severity-of-illness score designed for adult intensive care patients (Gray *et al,* 1992). This scoring system differs from the others reviewed here in two important ways. First, it uses the therapies received by the neonates rather than factors reflecting their condition. Second, the scoring system was derived from a panel comprising five neonatologists, a paediatric intensivist and a neonatal nurse as opposed to using the data to formulate a logistic regression model.

However, therapy depends on the practice and policy of units and can vary greatly even within a relatively small geographic area (Field *et al,* 2002). It is not possible to compare units using this type of adjustment. Such a system is likely to be of more use in the prediction of individual probabilities of death to aid counselling, for stratifying infants into risk groups in a trial or for helping to decide on treatment plans (Dorling *et al,* 2005).

4.4.4 National Institute of Child Health and Human Development (NICHHD)

Using a sample of 1,823 infants born between 1 November 1987 and 31 October 1989 this study was designed to develop a mortality risk model using admission perinatal factors for neonates weighing 501 to 1500 grams at birth (Horbar *et al,* 1993a). The infants were born at one of seven neonatal units in the USA.

The authors used a logistic regression model to select variables to be included in the model from a list of candidate variables known at the time of admission. The final model was then validated using a further 1,780 infants.

4.4.5 Berlin Score

This score aimed to quantify the mortality risk in infants with birth weights below 1500g (Maier *et al,* 1997). Derived using data on 572 infants, randomly split using 396 to develop the score and 176 for validation, the authors used logistic regression models to produce the final predictive model.

Although this score has the advantage that it only includes variables readily available at the time of admission, the inclusion of some of the variables is a cause for concern if the score is used to standardize between neonatal units (Tarnow-Mordi, 1997). The administration of

ventilatory support prior to admission to a neonatal unit is highly dependent on the policy of the hospital, and on the availability of equipment and facilities. Such policies and facilities are likely to vary greatly between obstetric units. There is also a difficulty with such heavy reliance on Apgar scores (discussed further in §4.4.8), and assessment of respiratory distress syndrome, as these allow an element of subjectivity that may lead to bias in any comparison between neonatal units.

A further difficulty with the Berlin score is that none of the infants included in this study had surfactant administered prior to admission. The scoring system should be reassessed as surfactant is now routinely administered to such infants, with a reported decrease in mortality (Horbar *et al,* 1993b; Horbar *et al,* 2002; Rosenberg *et al,* 2001; Suresh and Soll, 2001).

4.4.6 Sinkin Scores

Two scores were originally derived to predict the risk of bronchopulmonary dysplasia (defined as the need for supplemental oxygen at 28 days postnatal) amongst neonates at 12 hours and 10 days of life (Sinkin *et al,* 1990). However, it has been suggested that the scores *"are probably excellent mortality scores as well"* (Hentschel *et al,* 1998),

The scores were developed using data on 2341 infants admitted to one neonatal intensive care unit in New York. The 12-hour score could possibly be used for mortality risk-adjustment if it was found to work well. However, this suggestion was not supported by one study that used the Sinkin 12 hour score (SS_{12}) (Fleisher *et al,* 1997). Fleisher *et al* found that although the non-survivors had a greater observed mean score ($\mu = 2.25$: s.d. = 1.36) than the survivors ($\mu = 1.86$: s.d. = 1.06) the difference was not statistically significant ($p = 0.31$). However, this study lacked statistical power with only 10 non-survivors and 69 survivors included.

4.4.7 Neonatal Mortality Prognosis Index

In contrast to the scores described above that were derived for specific groups within the neonatal population, the Neonatal Mortality Prognostic Index was developed with the aim of predicting mortality before discharge for all infants admitted to neonatal intensive care (Garcia *et al,* 2000).

The score was derived 1994 using a logistic regression model with data from 336 infants (112 deaths) admitted to three neonatal units in Mexico City between July 1993 and August 1995. Potential prognostic factors were collected up to 12 hours after admission. The model was validated by examining sensitivity, specificity, positive and negative predictive values, using

an additional cohort of 300 infants (100 deaths). However, it is unclear if these were a random sample from all admissions or were a cohort from later years.

4.4.8 Apgar Score

The Apgar score was originally published in 1953 (Apgar, 1953) and is a simple, routinely used neonatal morbidity scoring system. It was designed to quantify a newborn infant's physical condition and to determine the level of care required. The score uses five features (heart rate, respiratory effort, muscle tone, reflex irritability and skin colour) determined by observation (Rennie and Roberton, 2002). The current practice is for infants to be scored twice: one and five minutes after birth (Letko, 1996). The early score aims to indicate the need for immediate treatment, whereas the second score indicates the infant's immediate response to resuscitation and the need for further intervention. Further Apgar scores may be calculated for an infant if its condition is felt to warrant it. The final score is on the scale 0 to 10, with higher scores representing healthier infants (see Appendix B).

Although not designed as a mortality score, an association with 28-day mortality was recognised in Virginia Apgar's original paper. Even though it is some 50 years since the creation of the score, there is still evidence for an association between low Apgar score and increase mortality, including in preterm infants (Casey *et al,* 2001: Weinberger *et al,* 2000).

Its simplicity means that the score has the potential to be unreliable for case-mix adjustment, at least on its own. Two studies have suggested that the accuracy of the score suffers due to poor inter-rater reliability, Clark and Hakanson (1988) and Livingston (1990). However, only eight descriptive cases were used in the former study, which concluded that nurses, in particular, inaccurately assigned Apgar scores. The second study comprised 52 infants, although this included only eleven preterm infants. Such sparse evidence, from North America from at least 15 years ago, gives little indication on the current reliability of Apgar scoring in UK neonatal units nor on the size or direction of any errors. While it is recognised that such errors may exist these can be overcome by adequate training of clinical and nursing staff (Letko, 1996). In addition, Apgar is not reliant on gestational age and birth weight for prediction, unlike CRIB II, and may offer the ability to appropriately 'fine tune' the morbidity of infants of equivalent birth weight and gestational age.

4.4.9 Transport Risk Index of Physiologic Stability

The Transport Risk Index of Physiologic Stability (TRIPS) was derived as an instrument to help assess the care given to neonates transported to tertiary NICUs in Canada (Lee *et al,* 2001). Data were collected on 1723 infants from January 1996 to October 1997, comprising 71% of all infants eligible for inclusion.

Although this scoring system was designed to investigate mortality within seven days of being transported, the authors also looked at its ability to predict total mortality before discharge. They reported that the index was a better predictor of such mortality than gestational age, although this did not hold for infants born at 32 weeks or less gestational age: TRIPS $AROC = 0.72$; gestational age $AROC = 0.75$. However, they did find that the combination of TRIPS, gestational and other risk factors (antenatal steroids, sex, SGA, 5-minute Apgar, vaginal delivery) improved the discriminatory ability of the model: $AROC = 0.83$. The score still needs validating in a population of infants that includes inborn infants.

TRIPS is calculated using four physiologic items: temperature, respiratory status, systolic BP, response to noxious stimuli. The advantage of TRIPS over most other neonatal mortality risk scores is that the observed values of all of its components, and the other variables included above, can be collected before admission to a neonatal unit.

4.5 Comparison of neonatal scores

The neonatal mortality risk scores described above vary both in complexity and in the variables that they include. The conflict between the requirement of a score to be simple to use, and therefore perhaps more reliable, but at the same time quantify complex risks is well recognised:

> *"It appears to be impossible for a score to be simple and parsimonious and at the same time to be rich, robust, and dynamic."* (Richardson *et al,* 2001)

In principle, a score such as CRIB II has a natural appeal as it is based on only five routinely collected variables. One study found that the original CRIB score took 5 minutes to apply whereas SNAP, SNAP-PE and NTISS took some 20 to 30 minutes each (Bastos *et al,* 1997). However, the question arises whether it is sufficiently complex to be able to quantify the risk

of mortality. Although the Trent Neonatal Survey data used here do not allow a comparison of the scores (only Apgar and the original CRIB are available) other studies have looked at this.

A study in Portugal of 186 infants, with birth weights less than 1,500g or born at under 32 weeks gestational age, found similar values for the area under the ROC curve (see §6.3.1) for four scoring systems: CRIB (0.90), SNAP (0.88), SNAP-PE (0.88), NTISS (0.85) (Bastos *et al,* 1997). Another study compared CRIB, SNAP-PE and NICHHD using 552 infants with birth weights from 500 to 1499 grams, born and admitted to 8 neonatal units between October 1994 and February 1997 (Pollack *et al,* 2000). Only small differences in the estimated area under the ROC curve were found: CRIB (0.89), SNAP-PE (0.91), NICHHD (0.87). A study of 222 neonates in Finland found that CRIB had better calibration (A_{ROC} = 0.89) compared to SNAP (0.82) and SNAP-PE (0.79) (Rautonen *et al,* 1994), and a small study from Brazil, 102 infants below 1,500g birth weight with 32 deaths, showed similar estimated area under the ROC curve for SNAP-PE (0.93), SNAP-PE II (0.94) and CRIB (0.91) (Zardo and Procianoy, 2003). Therefore, despite their differences, there does not appear to be large differences in the discriminatory ability of these scores.

It is also unclear whether such scoring systems can be used in countries other than the one in which they were developed. Problems may arise for two reasons. First, the variables that are predictive for infant mortality may not be the same in different populations or the weights used may not fit the new population (Rowan *et al,* 1993a). Second, even if the populations are similar, users may be more accurate in using scores that they are more familiar with (Richardson *et al,* 1994; Iezzoni *et al,* 1995). Also of importance, but rarely discussed in relation to neonatal risk scores, is the inter- and intra-observed reliability of the scores. These are important issues as a score inconsistently recorded is of limited use, especially if such differences are associated with the unit of care. It is likely that the simpler scores would also be the most reliable. This, and the evidence above suggesting equal performance of the models, suggests that the appropriate choice of scoring method is for the simple scores, such as CRIB.

One further point is that these scores, and the discussion in this thesis, relate to the prediction for groups of infants and not for individual predictions. Different risk scores may give similar overall predictions but individual estimates may differ greatly (Iezzoni *et al,* 1996a). While systems to predict an outcome for a particular infant may be of clinical use (Dorling *et al,* 2005) there are many difficulties with such an approach (Ridley, 2002; Lemeshow *et al,*

1995). Although this is an important research area, it is not directly relevant to the issues in this thesis and, therefore, will not be pursued further.

4.6 Risk Adjustment in this Thesis

Study specific risk-adjustment scores

It is possible to use the data available to develop a model specific to those data. This may be necessary where information on variables used in the pre-existing scores are not available, or where it is felt that these scores are not suitable for the particular population being investigated. However, particularly in the former case, it should be remembered that the local model would not include variables that other researchers have shown evidence of being associated with mortality.

Risk-adjustment approach taken in this thesis

The Trent Neonatal Survey was designed to collect information to allow adjustment for different mortality prognoses using CRIB. However, the deficiencies in the original CRIB have been recognised and an updated version, CRIB II, proposed. The Trent Neonatal Survey did not collect information on temperature at admission to the NICU, one of the five variables used in CRIB II (although the collection of this information was started from the beginning 2004). In addition, the birth weight by gestational age component of the CRIB II score is adapted from published mortality tables using previous TNS data (Draper *et al,* 1999). More up-to-date data are now available; both the data used in this thesis and as more recently published mortality tables (Draper *et al,* 2003). The application of CRIB and CRIB II to the TNS data is investigated in §6.9.4.

There is, therefore, no up-to-date pre-existing neonatal mortality risk score that is suitable for these data, except for the Apgar score. However, the Apgar score alone is unlikely to offer adequate risk-adjustment (although this is examined in 0), and a model to allow for mortality risk will be derived using the data available. It is accepted that such an approach may produced a model only applicable to these data but it is felt that such an approach is acceptable since it is not a requirement that the adjustment method developed here is generalizable to other data. In any case no pre-existing risk adjustment method is felt to be suitable, or at least superior to building a local model.

Therefore, the TNS data will be used to produce a risk-adjustment model.

4.7 Chapter Summary

In this Chapter various mortality risk-adjustment scores were described that have been proposed as suitable for provider profiling. However, none of these published scores were suitable for the TNS data in this thesis, either because the necessary data were not available or because it was felt that the scores do not adequately describe current outcomes in NICUs (§4.6). The approach taken here was to develop a risk-adjustment model from the available data and this is reported in Chapter 6.

In the next Chapter methods of presenting the data will be explored. In order to make these comparisons more realistic, the outcomes will be adjusted for gestational age at birth. This is known to have a very strong association with neonatal mortality, discussed in §6.4.1. A more complex model will be derived in Chapter 6, but it is felt that the methods in the next Chapter can be better demonstrated and discussed using a simpler model.

Chapter 5: OUTCOME SUMMARY MEASURES

5.1 Chapter Overview

In §3.4.1 the odds of mortality for each unit were estimated using a logistic regression model (Table 3.2). However, this does not give any information on whether the reported outcome for a unit differs from the rest of the Region. Alternative summary statistics that may be useful are discussed, illustrated and compared in this Chapter.

In §5.3 the Odds Ratio is described and its use in provider profiling illustrated. Three different parameterizations of the neonatal units are compared and the use of deviation contrasts in the rest of the thesis justified. Difficulties in the clinical interpretation of Odds Ratios are discussed and the alternative approach of standardization is explored in §5.4. Direct and indirect standardization are described, with proposed methods of significance testing (§5.4.2) and effect estimation (§5.5 & §5.9) reviewed. In particular, the use of standardized outcome ratios is explored in §5.5 and the use of the Standardized Mortality Ratio (SMR) proposed for this thesis. Methods for estimating confidence intervals for the SMR are described, and a Bayesian approach developed, in §5.6, and their properties were investigated through a simulation study, described in §5.7. The standardized mortality difference is briefly discussed in §5.9 and the use of random effects models is considered in §5.10. The main conclusions for the Chapter are discussed in §5.11.

5.2 Chapter Introduction

Gestational age at birth was included in the models throughout this Chapter, as it is known to have a strong relationship with infant mortality (Draper *et al,* 1999), discussed in more detail in §6.4.1. This allowed more realistic examples. It also allowed the *BCa* bootstrap method to be used in §5.6, as otherwise all observations within each unit would have the same predicted probability of death. This would have led to each bootstrapped sample taking the same value and, therefore, showing no variation and providing no estimated confidence intervals.

Hence, the logistic regression model used in this Chapter was (from *3.6*):

$$\log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_G.gest_i + \mathbf{X_u}\boldsymbol{\beta_u} \qquad (5.1)$$

where: *gest* is the gestational age at birth, in completed weeks

$\mathbf{X_u}$ is the design matrix specifying the units of admission

$\boldsymbol{\beta}_U$ is the vector of coefficients for the units of admission

Various potentially useful forms of the matrix $\mathbf{X_u}$ are discussed in §5.3.1.

The assumption was made that there was a linear relationship between gestational age at birth and mortality for all units. The plausibility of this assumption is investigated in Chapter 6.

All of the approaches in this Chapter allow the difference in outcome between a unit and a reference population to be represented by a single summary statistic. However, reducing differences between complex and heterogeneous populations to a single summary statistic means that much information is lost. Units may perform differently with different groups of patients: for example, a Scottish study of 30-day mortality after discharge following acute myocardial infarction found that the hospitals with the highest mortality rates for patients aged 50 years were different from those with the highest rates for patients aged 70 years (Leyland and Boddy, 1998). It is quite possible for a single figure to hide these differences. Thus, while an extreme value for a single statistic is likely to indicate abnormal rates of outcome, an unremarkable value may be hiding more complex differences. In addition, the use of summary measures is of no use to individual patients who would want to know the probabilities specific to them (Rao, 2001).

Nevertheless, these are useful methods. Much of the role of statistical methods is to provide meaningful summaries of data to allow inferences to be drawn:

> *"Such summary statistics are ... necessary for the precise and efficient comparison of different sets of data."* (Hennekens and Buring, 1987:229)

There is a responsibility on the statistician and the user to ensure that appropriate conclusions are made from these summary statistics.

5.3 Odds Ratio

While the probability of death in individual units can be estimated with the methods discussed in §3.4.1, this says nothing about the relative sizes of such probabilities compare to the rest of the Region. One method to investigate the relative sizes of the mortality rates is to compare the odds of death for infants in a unit to that in the rest of the Region. Since odds are additive on the logarithmic scale, it is, in general, the ratio of two odds that is most often compared. This can be done in different ways. In particular, the method for the parameterisation of the health care providers within the logistic regression model will depend on the way the results are to be reported.

The standard logistic modelling approach is to choose one of the providers as a reference; sometimes called **reference cell coding** (Hosmer and Lemeshow, 2000:56). For example, in a comparison of infant mortality between Canadian provinces and territories, Quebec was chosen as the reference as it was the largest care provider and also had the lowest mortality rate (Wen *et al,* 2000). However, the most obvious problem with this approach is finding the most appropriate choice of reference. Sometimes there may be a provider that naturally fits the role of reference. If not, another way, that used in the example above, is to choose the 'best' performing provider and then test whether there is evidence that the other units are performing poorly compared to this 'best' provider. However, there are difficulties with this. First, the reference cannot be identified *a priori*, rather it is determined by the data. Second, the choice of the reference provider is arbitrary and may not be relevant organisationally: for example, the reference may be a small unit with unusual working practices. It is also the case that the 'best' performing provider may be difficult to identify: for example, if the outcome of interest is the use of mechanical ventilation it is likely that neither extremely high nor extremely low rates are 'best' practice. In addition, in this thesis, the interest is primarily in comparing each unit to the rest of the Region, rather than to one particular unit.

While in some cases it may be possible to identify a suitable reference provider *a priori*, alternative, perhaps more useful, approaches exist. These will be examined in the following sub-section.

5.3.1 Parameterization of the reference units

Rest of Region

The simplest approach is to compare the outcome in the unit of interest to the rest of the Region as a single group. Hence, the hypotheses of interest are:

$$H_0 : g_j = g_{R(-j)}$$

$$H_1 : g_j \neq g_{R(-j)}$$

where: g_j is the log odds for the provider of interest;

$g_{R(-j)}$ is the log odds for the Region excluding observations from the provider of interest.

The logistic regression model can be written, from (*5.1*):

$$log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_G.gest_i + \beta_j.I_{ji} \qquad (5.2)$$

where: $I_j = \begin{cases} 1 & \text{if admitted to Unit j} \\ 0 & \text{otherwise} \end{cases}$

Therefore, the natural logarithm of the odds ratio of interest (ψ_j) is:

$$\log_e(\psi_j) = g_j - g_{R(-j)} = \beta_j$$

Table 5.1 shows the results of such analyses using the TNS data.

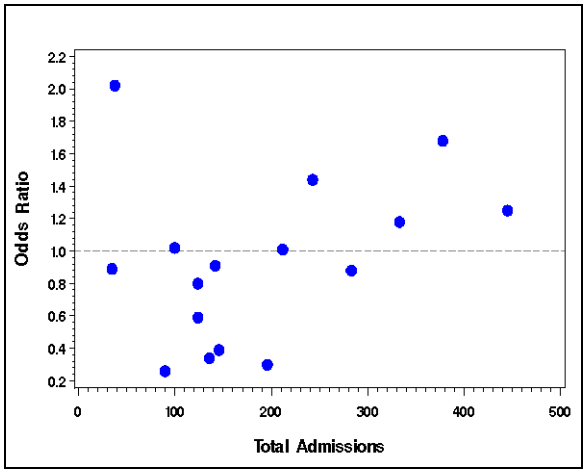
The estimated odds ratios varied greatly, from 0.26 (Unit 14) to 2.02 (Unit 3). However, both of these values had very wide associated 95% confidence intervals. Two units showed evidence, at the 5% significance level, of extreme odds ratios: Unit 6 had a high mortality rate and Unit 12 had a low rate.

Table 5.1 Odds ratios for in-unit mortality with Wald confidence intervals

Unit	Odds Ratio	Wald 95% CI	Wald p-value
1	1.01	(0.56 to 1.78)	0.99
2	0.88	(0.54 to 1.43)	0.61
3	2.02	(0.33 to 12.10)	0.44
4	0.91	(0.37 to 2.25)	0.84
5	1.18	(0.77 to 1.78)	0.44
6	1.68	(1.13 to 2.48)	0.0094
7	1.44	(0.87 to 2.36)	0.15
8	0.59	(0.25 to 1.38)	0.22
9	0.89	(0.10 to 7.37)	0.91
10	0.39	(0.14 to 1.04)	0.058
11	1.25	(0.87 to 1.80)	0.22
12	0.30	(0.11 to 0.77)	0.012
13	0.34	(0.09 to 1.22)	0.097
14	0.26	(0.06 to 1.16)	0.077
15	0.80	(0.37 to 1.71)	0.56
16	1.02	(0.40 to 2.57)	0.97

However, even after allowing for the gestational age of infants at birth, an association still remained between the number of admissions to a unit and its rate of mortality (Figure 5.1). It is uncertain whether this is due to the quality of care or to differences in the morbidity of the infants. This will be investigated further in Chapter 6.

Figure 5.1 Adjusted odds ratios by total admissions



What is clear is that the larger units can influence the previous results in two ways. First, collectively they raise the estimate of the 'rest-of-region' odds ratio ($g_{R(-j)}$). Second, they influence the estimate for the effect of gestational age (β_G).

The use of a common parameter estimate for the relationship between gestational age and mortality is not a problem. Using this model specification it is assumed that there is a constant difference (on the logit scale) between the units in gestational age specific odds of death. However, the extent to which large units influence the estimation of $g_{R(-j)}$ is a problem. In this example, a large outlying unit is less likely to be identified whereas small units are more likely. Approaches to overcome this problem are discussed below.

Weighted Logistic Regression

The problem of large units erroneously influencing the estimation odds of death for 'the rest of the Region' can be overcome by using weighted logistic regression. The contribution of each observation to the log likelihood is given a weight (w_i):

$$\text{Log L} = \sum_{i=1}^{N} w_i \log(\hat{\pi}_i)$$

The simplest way of assigning weights is to use $\frac{1}{n_j}$ as the weight for each observation in Unit j: where n_j is the number of observations in Unit j. To ensure that the estimated covariance matrix is invariant to the weights, the weights are normalized so that they sum to the actual sample size. For the weights used here, each weight is multiplied by the mean number of admissions for the units: i.e. for observations in unit j their normalized weight is $\frac{\sum_{k=1}^{16} n_k}{16} \Big/ n_j$.

This can be achieved directly in SAS `PROC LOGISTIC` using the `NORMALIZE` option in the `WEIGHT` statement (SAS Institute Inc., 1999:1939):

i.e.
```
weight wt / normalize;
```

However, intuitively the down-weighting (or up-weighting) of the unit of interest seems inappropriate. This would overestimate the variance for large units and underestimate that of a small unit. The approach to be taken here is for the 15 reference units to be weighted to ensure equal importance and for each observation in the unit of interest take the weight value 1. For observations from each of the reference units the weight is:

$$\frac{\sum_{\substack{k=1 \\ k \neq j}}^{16} n_k}{16} \Big/ n_j \qquad (5.3)$$

The model remains:

$$\log_e(\pi) = \beta_0 + \beta_G.gest + \beta_j.I_j$$

where: $I_j = \begin{cases} 1 & \text{if admitted to Unit j} \\ 0 & \text{otherwise} \end{cases}$

The natural logarithm of the odds ratio of interest *(ψ_j)* is still:

$$\log_e(\psi_I) = g_j - g_{R(-j)} = \beta_j$$

As before, the indicator variable can be created in a SAS `DATASTEP` and then included in the `MODEL` (and `CLASS`) statement in `PROC LOGISTIC`. The weights can also be created using a SAS `DATASTEP` and then specified in the `WEIGHT` statement in `PROC LOGISTIC`. The SAS macro ***weighted*** that does this is shown in Appendix D.2.

The results of these analyses are shown in Table 5.2. There was statistical evidence that Units 6 and 12 had extreme odds ratios.

The estimated odds ratios were higher in the weighted model than in the unweighted (Table 5.1 and Table 5.2). This was because reducing the influence of the large (high mortality) units reduced the mortality rate in the reference group for each comparison.

However, by using weighted logistic regression the parameter estimates for risk factors became 'weighted' estimates. Weighted estimates are less efficient than estimates from unweighted models (Harrell, 2001:206). An alternative, more efficient, approach is to use deviation contrasts; such contrasts are discussed next.

Table 5.2 Weighted logistic regression

Unit	Odds Ratio	(95% Confidence Interval)	P-value
1	1.14	(0.63 to 2.05)	0.66
2	1.00	(0.61 to 1.64)	0.99
3	2.47	(0.40 to 15.14)	0.33
4	1.07	(0.42 to 2.66)	0.89
5	1.33	(0.87 to 2.04)	0.18
6	1.86	(1.24 to 2.77)	0.024
7	1.64	(0.98 to 2.73)	0.056
8	0.65	(0.27 to 1.54)	0.32
9	1.03	(0.12 to 8.63)	0.98
10	0.44	(0.16 to 1.17)	0.10
11	1.39	(0.95 to 2.03)	0.082
12	0.33	(0.12 to 0.87)	0.025
13	0.39	(0.10 to 1.40)	0.15
14	0.29	(0.06 to 1.28)	0.10
15	0.90	(0.41 to 1.94)	0.79
16	1.19	(0.46 to 3.04)	0.72

Deviation from the mean

An alternative approach is to compare the log odds of the provider of interest to the mean of the estimated log odds of the reference units, for example Sankaran *et al* (2002). These are referred to as **effect** contrasts within SAS (SAS Institute Inc., 1999:1915), but are also called **deviation**, or **deviation from means**, contrasts (Hosmer and Lemeshow, 2000:54).

In the standard application of these contrasts, the hypotheses of interest are:

$$H_0 : g_j = \frac{\sum_{k=1}^{N} g_k}{N}$$

$$H_1 : g_j \neq \frac{\sum_{k=1}^{N} g_k}{N}$$

where: g_k is the log odds for the outcome in unit k: $k \in \{1,2,\ldots,16\}$;

g_j is the log odds for the provider of interest;

N is the total number of providers.

Therefore, the natural logarithm of the odds ratio of interest (ψ_i) is:

$$\log_e(\psi_j) = g_j - \frac{\left(\sum_{k=1}^{N} g_k\right)}{N}$$

$$= \frac{(N-1)}{N} g_j - \frac{g_1}{N} - \frac{g_2}{N} - \ldots - \frac{g_{N(-j)}}{N}$$

In this case the model is:

$$logit(\pi) = \beta_0 + \beta_G.gest + \beta_1.I_1 + \beta_2.I_2 + \ldots + \beta_{15}.I_{15} \qquad (5.4)$$

where:
$$\left.\begin{array}{l} I_j = 1: \text{ if admitted to Unit j} \\ I_j = -1: \text{if admitted to Unit 16} \\ I_j = 0: \text{ otherwise} \end{array}\right\} \quad j \in \{1,2,\ldots,15\}$$

From (*5.4*) it can be seen that the log odds of death before discharge are:

for Units $j = \{1, 2,\ldots,15\}$: $\text{logit}(P_D \mid gestation) = \beta_0 + \beta_G.gest + \beta_j$

and for Unit 16: $\text{logit}(P_D \mid gestation) = \beta_0 + \beta_G.gest - \sum_{k=1}^{15}\beta_k$

Furthermore, it is simple to show that, when $gest = 0$, the intercept (β_0) takes the value of the mean of the estimated log odds $\left(\frac{\sum_{k=1}^{16} g_k}{16}\right)$:

$$\frac{\sum_{k=1}^{16} g_k}{16} = \frac{1}{16}\left(\left[\beta_0 + \beta_G.gest + \beta_1\right] + \ldots + \left[\beta_0 + \beta_G.gest + \beta_{15}\right] + \left[\beta_0 + \beta_G.gest - \sum_{k=1}^{15}\beta_k\right]\right)$$

$$= \frac{1}{16}\left(16.\beta_0 + 16.\beta_G.gest + \left[\beta_1 + \beta_2 + \ldots + \beta_{15} - \sum_{k=1}^{15}\beta_k\right]\right)$$

$$= \beta_0 + \beta_G.gest$$

Hence, using the parameterisation in (*5.4*), the estimated regression coefficient for Units 1 to 15 (β_1 to β_{15}) represents the estimated log odds ratio for that unit compared to the mean outcome. For Unit 16 the corresponding log odds ratio is estimated by $\left(-\sum_{k=1}^{15}\beta_k\right)$.

Although each unit is compared to the mean of the log odds, each unit still has an influence on the mean to which it is being compared: for example, a unit with a high mortality rate will increase the value of the observed mean log odds. This produces conservative results since each unit influences the average towards its own observed value. An alternative approach to overcome this problem would be to estimate the mean regional response based on the other fifteen units and then compare the unit of interest to that value. This approach most closely matches that in the 'rest of Region' and 'weighted' models, in that the unit of interest is compared to some 'average' odds for the rest of the Region.

In this case the hypotheses of interest becomes:

$$H_0 : g_j = \frac{\left(\sum_{k=1}^{16} g_k\right) - g_j}{15} = \frac{\left(\sum_{\substack{k=1 \\ k\neq j}}^{16} g_k\right)}{15}$$

$$H_1 : g_j \neq \frac{\left(\sum_{k=1}^{16} g_k\right) - g_j}{15} = \frac{\left(\sum_{\substack{k=1 \\ k\neq j}}^{16} g_k\right)}{15}$$

The log odds ratio of interest is now given by:

$$Log_e\,(Odds\ Ratio) \qquad = g_j - \frac{\left(\sum_{\substack{k=1 \\ k\neq j}}^{16} g_k\right)}{15}$$

Such estimates can be obtained from SAS `PROC LOGISTIC` by using appropriate `CLASS` and `CONTRAST` statements and, in this example, using the `EFFECT` indicator contrasts with Unit 16 as the reference category (i.e. `class c_hosp / param=effect ref=last;`).

Using Unit 1 as an example, the log odds of death are:

$$\left(LogOdds_{(1)} \mid gest\right) = \beta_0 + \beta_G.gest + \beta_1 \qquad (5.5)$$

However, the mean log odds for the other units is now given by:

$$\begin{aligned}
&\left(LogOdds_{(2-16)} \mid gest\right) \\
&= \tfrac{1}{15}\left(\left[\beta_0 + \beta_G.gest + \beta_2\right] + \ldots + \left[\beta_0 + \beta_G.gest + \beta_{15}\right] + \left[\beta_0 + \beta_G.gest - \sum_{j=1}^{15}\beta_j\right]\right) \\
&= \tfrac{1}{15}\left(15.\beta_0 + 15.\beta_G.gest + \left[\beta_2 + \beta_3 + \ldots + \beta_{15} - \sum_{j=1}^{15}\beta_j\right]\right) \\
&= \beta_0 + \beta_G.gest - \frac{\beta_1}{15} \qquad (5.6)
\end{aligned}$$

Therefore, the test of the log odds of Unit 1 versus the mean of the log odds of the other units is:

$$\begin{aligned}
\beta_0 + \beta_G.gest + \beta_1 &= \beta_0 + \beta_G.gest - \tfrac{\beta_1}{15} \\
0 &= \beta_1 + \tfrac{\beta_1}{15} \\
0 &= \tfrac{16}{15}\beta_1
\end{aligned}$$

This is straightforward to test using a `CONTRAST` statement in `PROC LOGISTIC`:

i.e.
```
contrast 'unit 1'
      c_hosp  1.0667 0 0 0 0 0 0 0 0 0 0 0 0 0 0 /estimate=exp;
```

(Since $\frac{16}{15} \approx 1.0667$)

Appropriately modified `CONTRAST` statements can be used for Units 2 to 15 and it can be shown that the equivalent test for Unit 16 is:

$$0 = -\tfrac{16}{15}\sum_{j=1}^{15}\beta_j$$

i.e.
```
contrast 'unit 16' c_hosp     -1.0667 -1.0667 -1.0667 -1.0667
                              -1.0667 -1.0667 -1.0667 -1.0667
                              -1.0667 -1.0667 -1.0667 -1.0667
                              -1.0667 -1.0667 -1.0667 / estimate=exp;
```

The results from these analyses are shown in Table 5.3.

Table 5.3 Deviation contrast Odds Ratios

Unit	Odds Ratio	(95% Confidence Interval)	P-value
1	1.29	(0.70 to 2.39)	0.41
2	1.14	(0.67 to 1.95)	0.62
3	2.62	(0.43 to 15.58)	0.29
4	1.15	(0.45 to 2.92)	0.76
5	1.49	(0.93 to 2.39)	0.095
6	2.05	(1.30 to 3.21)	0.0017
7	1.83	(1.05 to 3.15)	0.030
8	0.75	(0.31 to 1.80)	0.52
9	1.11	(0.13 to 9.23)	0.92
10	0.49	(0.17 to 1.32)	0.16
11	1.56	(1.02 to 2.39)	0.040
12	0.37	(0.13 to 0.98)	0.044
13	0.42	(0.11 to 1.51)	0.18
14	0.32	(0.07 to 1.41)	0.13
15	1.02	(0.46 to 2.25)	0.96
16	1.29	(0.50 to 3.34)	0.59

The estimated values for the odds ratios are now greater than those from the previous models. The results from all three models are discussed next.

Comparison of parameterizations

The three methods (unweighted and weighted 'rest of Region' parameterization, and 'deviation' parameterisation) outlined above produce three different solutions, as illustrated by the following example where Unit 1 is to be compared to the combined outcome of Units 2 to 16. When there is no risk adjustment the combined outcomes are given in Table 5.4.

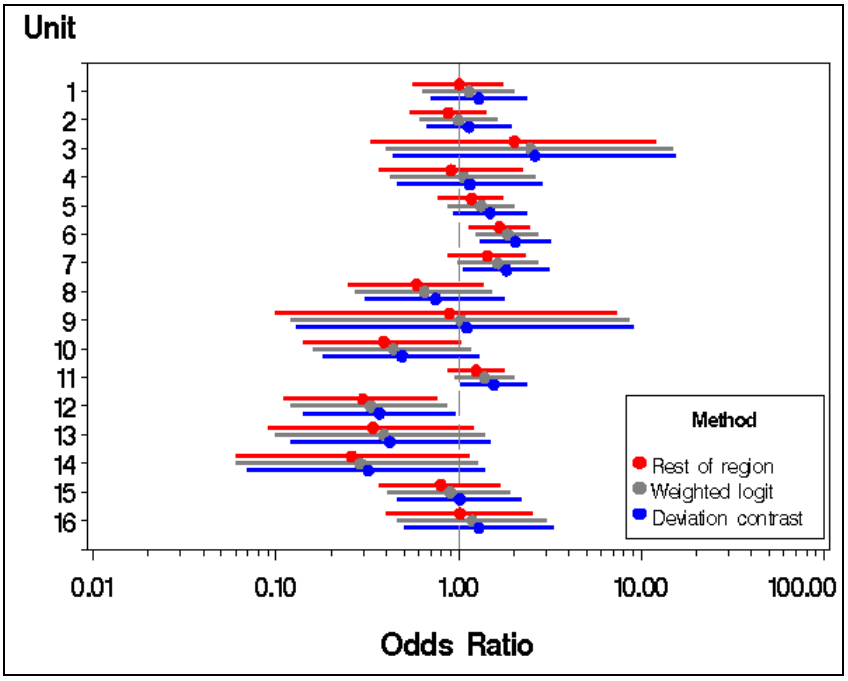
Hence, in an unadjusted analysis, the odds of death in Unit 1 ($\pi_1 = 0.0991$; $\omega_1 = 0.1099$) are compared to three very different values. Such differences also hold when the model contains additional variables to allow for morbidity differences, in this Chapter gestational age at birth. However, the different models also result in different estimates for the association between gestational age and mortality due to the different weights given to the units.

Table 5.4 Unadjusted summary Odds (Units 2 to 16) to which Unit 1 is compared

Method	Odds (ω)	$\hat{\omega}$	$\hat{\pi}$
Rest of region:	$\omega_{(2-16)} = \dfrac{\left(\dfrac{\sum_{j=2}^{16}\sum_{i=1}^{n_j}\pi_{ij}}{\sum_{j=1}^{16} n_j}\right)}{1-\left(\dfrac{\sum_{j=2}^{16}\sum_{i=1}^{n_j}\pi_{ij}}{\sum_{j=1}^{16} n_j}\right)}$	0.1036	0.0939
Weighted logistic regression:	$\omega_{(2-16)} = \dfrac{\left(\dfrac{\sum_{j=2}^{16}\left[\frac{\sum_{i=1}^{n_j}\pi_i}{n_j}\right]}{15}\right)}{1-\left(\dfrac{\sum_{j=2}^{16}\left[\frac{\sum_{i=1}^{n_j}\pi_i}{n_j}\right]}{15}\right)}$	0.0763	0.0709
Deviation contrasts:	$\omega_{(2-16)} = \exp\left(\dfrac{\sum_{j=2}^{16} g_j}{15}\right)$	0.0622	0.0585

The differences in the odds ratios can be seen in Figure 5.2, where the odds ratios estimated using the first method have a lower absolute value than the other two methods. Although the estimates and their 95% confidence intervals are similar there are different conclusions that can be drawn at the 5% significance level for units 7 and 11.

Figure 5.2 Odds Ratios estimated using three different methods



The method felt to be most appropriate with this thesis is **deviation contrasts**. There are two reasons for this. First, the 'rest of region' method gives unjustifiable influence to the larger units. Second, using weighted logistic regression causes the estimates for additional risk factors, in this case gestational age, to be 'weighted' estimates and, as outlined previously, weighted estimates are less efficient than estimates from unweighted models (Harrell, 2001:206). The use of deviance contrasts overcomes both of these problems.

5.3.2 Bayesian Analysis

The use of a Bayesian approach allows the assumption of a probability distribution for the odds ratios, based on the underlying binomial distributions. This has the advantage of allowing probability statements to be made regarding them: for example, what is the probability the odds ratio is greater than 2? Such questions allow the emphasis to rest with clinical significance rather than purely statistical significance (Burton *et al,* 1998), as the probability that a provider's performance is extreme (i.e. an outlier) is not the same as saying that they have a clinically extreme mortality rate. For large providers small differences in performance may be statistically significant, so instances of classifying providers as 'low-performers' purely on the basis of a statistical significance is unwise (Glance *et al,* 2002). Clinical standards are likely to be of most use when selected by those who are going to use the results but it is possible, indeed likely, that these different users will have different expectations of appropriate benchmarks (Christiansen and Morris, 1997). However, if the

posterior distribution can be specified then the different posterior probabilities can be easily calculated.

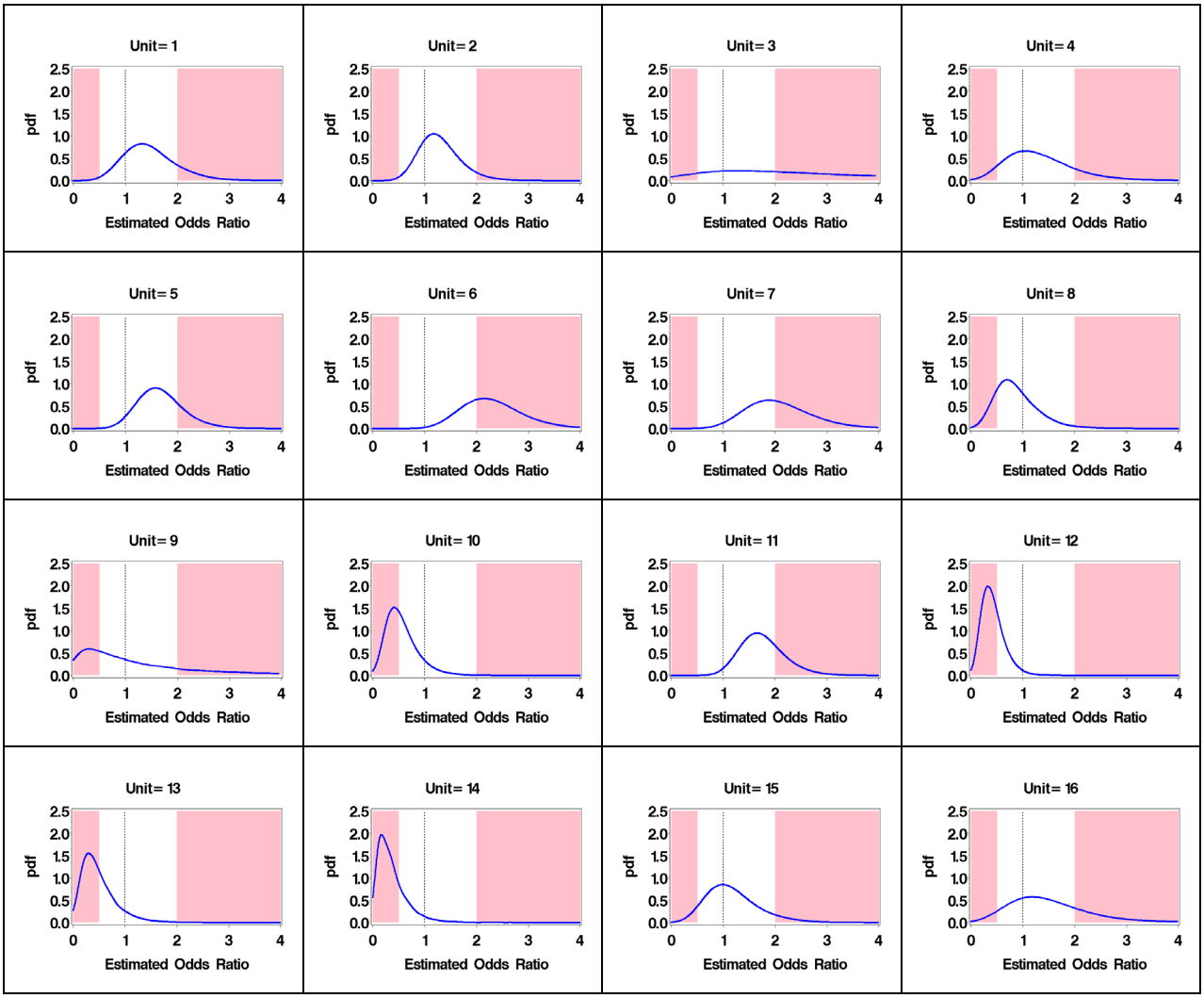
A Bayesian approach is illustrated using WinBUGS (§3.2.2) to estimate parameters for the deviance model described previously (§5.3.1). The outcomes were estimated using $N(0,1000^2)$ as the prior distribution for all of the parameter estimates. A 1,000-iteration burn-in was inspected using five independent chains and Brooks-Gelman-Rubin statistics calculated for the model parameters. There was no evidence that the model exhibited poor mixing and the parameters were estimated using a further 10,000 sampled values (Appendix E.2). In addition, the estimated odds ratio was inspected at each iteration and the proportion of times its value fell below 0.5 ($P(\psi<1/2)$), or above 2 ($P(\psi>2)$), recorded. The results are shown in Table 5.5.

Table 5.5 Odds ratio estimated using Bayesian approach

Unit	Odds Ratio	(95% Credible Interval)	$P(\psi<1/2)$	$P(\psi>2)$
1	1.45	(0.73 to 2.82)	0.0010	0.17
2	1.29	(0.73 to 2.27)	0.0011	0.060
3	2.50	(0.25 to 12.14)	0.075	0.59
4	1.25	(0.43 to 3.07)	0.042	0.16
5	1.69	(1.03 to 2.78)	<0.0001	0.25
6	2.31	(1.45 to 3.79)	<0.0001	0.73
7	2.06	(1.17 to 3.64)	<0.0001	0.54
8	0.82	(0.3 to 1.94)	0.15	0.022
9	0.88	(0.02 to 5.66)	0.34	0.23
10	0.52	(0.16 to 1.32)	0.47	0.0023
11	1.77	(1.12 to 2.84)	<0.0001	0.30
12	0.39	(0.12 to 0.96)	0.67	0.0001
13	0.43	(0.08 to 1.37)	0.59	0.0034
14	0.31	(0.04 to 1.2)	0.74	0.0025
15	1.12	(0.46 to 2.5)	0.036	0.076
16	1.38	(0.46 to 3.43)	0.031	0.22

The smoothed posterior probability density functions are also shown (Figure 5.3) to further illustrate the interpretation of $P(\psi<1/2)$ and $P(\psi>2)$.

Figure 5.3 Estimated posterior probablilty density functions for Odds Ratio



The Bayesian estimates in Table 5.5 can be compared with the estimates in Table 5.3. When the point estimates are compared, differences can be seen. For all units except four (Units 3, 9, 14 and 16) the estimates for the odds ratio are higher in the Bayesian analysis than from the classical model. The four units where this is not the case are the four smallest units in the study. This difference between the two approaches is due to the introduction of prior probability distributions for the model parameters. When the units are small, the information from the prior probability dominates that from the data. In this case, the prior probability distribution chosen for all of the parameter estimates was $N(0,1000^2)$. Although this distribution may be 'vague', that is having a large variance, it still has a mean higher that the observed values for β_1, β_2,…, β_{15}. Where there is little information in the data the model estimates for these parameters are drawn ('shrunk') toward the mean of the prior distribution. Due to the model parameterization this affects all of the parameter estimates. One solution lies with a more considered choice of prior distributions, as was discussed in §3.4.2, although

with such sparse data, any prior distribution will be highly influential for the small units. This is pursued further in §5.8.4.

5.3.3 Use of odds ratios

This section has outlined approaches that use the odds ratio to summarize the difference between a unit and the rest of the Region. However, odds ratios can be difficult to interpret and, in practice, they are often interpreted as relative risks, although they can be poor approximations (Sinclair and Bracken, 1994; Davies *et al,* 1998; Sackett *et al,* 1996). One solution may be to use a generalized linear model to estimate relative risks directly, either by using a log-log link (Martuzzi and Elliott, 1998) or a log-link (Wacholder, 1986), as these have a straightforward interpretation. However, such models present difficulties, in particular problems with convergence, but also predicted probabilities outside of the interval [0,1] and confidence intervals that are too small (McNutt *et al,* 2003; Wacholder, 1986; Martuzzi and Elliott, 1998).

Although possible solutions may exist to these problems, alternative summary statistics have been suggested using logistic regression models. These are commonly used in practice and are explored in the next Section.

5.4 Standardization

5.4.1 Direct and indirect standardization

The aim of **standardization** is to provide a measure of the difference between the population of interest and a standard, or reference, population. To do this, the sample of interest and the reference population are divided into relatively homogeneous strata (risk-adjustment) and the stratum specific mortality rates of the two data sets calculated and compared. Possible summary statistics that can be used are illustrated later in this Chapter, but before that two approaches to standardization, **direct** and **indirect**, are discussed and compared.

Direct standardization

Direct standardization provides an answer to the question, 'What would have been the outcome in the rest of the Region if its population had the same outcomes as those in the unit of interest?' To do this, stratum specific event rates are calculated for the population of

interest and these rates are then applied to the reference population and the expected number of events ($\boldsymbol{D_{DIRECT}}$) calculated:

$$D_{DIRECT} = \sum_{i=1}^{l} \pi_i n_{Ri} \qquad (5.7)$$

where: l is the number of strata

π_i is the probability of death for an observation in stratum i in unit of interest

n_{Ri} is the number of observations in stratum i in the reference population

Alternatively, this can be written as:

$$D_{DIRECT} = N_R \sum_{i=1}^{l} \pi_i p_{Ri} \qquad (5.8)$$

where: N_R is the total number of observations in the reference population

p_{Ri} is the proportion of observations in the reference population in stratum i

Since the expected number of deaths is dependent on the population structure of the reference population, its value, and that of any summary statistic derived from it, is difficult to interpret in relation to the unit of interest. On the other hand, if various units are standardized to the same reference population then it is appropriate to use the values obtained to compare the units with each other. This property will be discussed in more detail later (§5.5.1).

The main problem with using direct standardization is that the estimation of stratum specific mortality rates for the units of interest can be difficult. These units are often small, as with the TNS data, and any estimated rates are likely to have large sampling errors. It seems inconceivable that the small volume neonatal units have a sufficient number of observations to allow the creation of sufficiently homogeneous strata. In fact, inspection of Table 2.5 shows that the estimated mortality rates for the whole sample of each unit have a large amount of uncertainty. At the extreme, Unit 9 has only one death from 35 admissions, giving an exact 95% confidence interval for the true death rate from 0.000 to 0.150, and once there is more than one stratum all but one will have an observed death rate of zero. While this example is at the extreme, a similar argument exists for all but the very largest units. Even if direct standardization was possible by using only a small number of strata, the mortality rate estimates are still likely be poorly estimated.

However, it may be possible to obtain directly standardized outcomes by including a risk-adjustment covariate in the model as a continuous variable, rather than as a categorical

variable. To illustrate this, directly standardized expected mortality is shown in Table 5.6 for the TNS data in this thesis. Two expected mortality totals are shown; first without adjustment for gestational age (i.e. using the overall mortality rate for each unit) and then directly standardized using gestational age at birth as a continuous variable in a logistic regression model.

When investigating Unit *j*, the model for estimating the probability of death for observation *i*, adjusted for gestational age, is:

$$logit(\pi_i) = g_i = \beta_0 + \beta_G . gest_i \qquad (5.9)$$

Hence, the estimated probability of death for an individual in the reference population is:

$$\hat{\pi}_{Ri} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_G . gest_i)}} \qquad (5.10)$$

and the expected number of deaths in the rest of the Regional (D_{DIRECT}) is given by summing these estimated probabilities for all observations across all fifteen reference units:

$$D_{DIRECT} = \sum_{\substack{k=1 \\ k \neq j}}^{16} \sum_{i=1}^{n_k} \hat{\pi}_{Ri} \qquad (5.11)$$

The expected number of deaths from direct standardization adjusted for gestational age is generally closer to the observed number than the unadjusted expected number of deaths. This suggests that differences in the gestational ages of the infants may account for some of the differences in mortality rates between the units. However, uncertainty in the estimates of β_0 and β_G mean that uncertainty around D_{DIRECT} is likely to be large. For small units, for example Units 3 and 9, the difference is very large, suggesting that the effect of gestational age on mortality may be poorly estimated.

Table 5.6 Observed and expected number of deaths: directly standardized for gestational age at birth

Unit	Observed (Σd_i)	Died	
		Unadjusted standardized expected ($\Sigma \pi_i$)	Adjusted standardized expected ($\Sigma \pi_i$)
1	264	278.6	260.3
2	255	290.7	244.0
3	283	157.2	822.0
4	279	120.9	295.1
5	244	331.4	269.1
6	231	378.1	324.4
7	256	332.0	322.9
8	277	187.2	194.4
9	284	85.4	379.5
10	280	98.6	148.2
11	223	360.4	259.5
12	280	72.2	96.8
13	282	63.7	139.2
14	283	65.2	95.1
15	275	234.0	240.6
16	279	175.5	263.6

Indirect standardization

To overcome the problem seen with direct standardization of trying to estimate stratum specific mortality rates from the unit of interest, **indirect standardization** applies the mortality rates observed in the reference population to the population of interest. This amounts to the general question, 'What would have been the outcome if the Unit's population had the same outcome as patients with the same characteristics in the rest of the Region?' This is likely to overcome the problem of imprecise stratum specific rates seen with direct standardization, as the reference population is usually larger and has a greater number of observations in each stratum.

Following on from (*5.7*) *&* (*5.8*), indirect standardization can be written:

$$D_{INDIRECT} = \sum_{i=1}^{l} \pi_{Ri} n_i \qquad (5.12)$$

where: l is the number of strata

π_{Ri} is the probability of death in reference population for an observation in stratum i

n_i is the number of observations in stratum i in the unit of interest

Or as: $$D_{INDIRECT} = N \sum_{i=1}^{l} \pi_{Ri} p_i \qquad (5.13)$$

where: N is the total number of observations in the unit of interest

p_i is the proportion of observations in the unit of interest in stratum i

Using indirect standardization, the expected number of deaths for each provider is weighted according to the characteristics of their population, i.e. p_1, p_2,…,p_l (Bhopal, 2002:194-198). Since, in many practical situations, each unit is likely to have samples with different empirical distributions of the variables used for standardization (risk adjustment) then the rates estimated for each unit will, strictly, not be comparable with each other. This has been known for a long time:

> *"… it will appear evident that if any one locality had an excess of population at that period* [age group] *where the mortality was 25 per cent., and a deficiency of population at that period at that period where the rate of mortality was only half per cent., that the average amount of mortality, the number of deaths … would differ widely from that of another locality in which the order of population was exactly reversed …"* (Neison, 1844)

This characteristic will just be noted here but will be addressed further in §5.5.

In the meantime, the TNS data were used to indirectly standardize each neonatal unit to the rest of the Region using gestational age at birth as a continuous variable in a logistic regression model. The model used in this section is slightly different to that used in §5.3.1, in that the data from the unit of interest were not used in the model to find the Regional average. The main difference is that the estimate for the effect of gestational age is now estimated from the 15 units that comprise the 'rest of Region' rather that all 16 units as previously. In practice, this has almost no effect on the parameter estimates and, therefore, the expected number of deaths for each unit. The main disadvantage with this second approach is that 16

separate models are now required rather than the previous single model. However, the advantage gained is that any estimates from the two parts of the data ('unit of interest' & 'rest of Region') are statistically independent. This characteristic will be utilized in §5.6.

So, for example, when investigating Unit 1 the model can be written as:

$$logit(\pi_i) = g_i = \beta_0 + \beta_G.gest_i + \sum_{k=2}^{16} \beta_k I_k \qquad (5.14)$$

where:
$$\left.\begin{array}{l} I_k = 1: \quad \text{if admitted to Unit k} \\ I_k = -1: \quad \text{if admitted to Unit 16} \\ I_k = 0: \quad \text{otherwise} \end{array}\right\} \quad k \in \{2,...,15\}$$

Hence, the estimated probability of death for an individual in Unit 1 is:

$$\hat{\pi}_i = \frac{1}{1 + e^{-\left(\hat{\beta}_0 + \hat{\beta}_G.gest_i\right)}} \qquad (5.15)$$

Summing these gives the expected total deaths in Unit 1, $D_{1INDIRECT}$:

$$D_{INDIRECT} = \sum_{i=1}^{n_1} \hat{\pi}_i \qquad (5.16)$$

This process was carried out for each unit both with and without adjustment for gestational age at birth (Table 5.7).

For most of the units (the exceptions are Units 3, 8 & 16) the standardized expected number of deaths is closer in value to the observed number of deaths than the unstandardized. As before, this suggests that there are differences in the gestational age structures between the units and that these differences account for some of the differences in the observed mortality rates.

Table 5.7 Observed and expected number of deaths: indirectly standardized for gestational age at birth

Unit	Died		
	Observed (Σd_i)	Unadjusted standardized expected ($\Sigma \pi_i$)	Adjusted standardized expected ($\Sigma \pi_i$)
1	21	12.4	17.8
2	30	16.5	27.9
3	2	2.3	1.1
4	6	8.9	5.3
5	41	19.2	31.1
6	54	21.6	35.1
7	29	14.0	19.7
8	8	7.5	9.8
9	1	2.2	0.9
10	5	9.2	8.9
11	62	25.4	46.9
12	5	12.6	11.5
13	3	8.8	6.0
14	2	5.8	5.2
15	10	7.4	9.9
16	6	6.0	4.9

Use of direct and indirect standardization

Although direct standardization has been used in published research, for example Horbar *et al* (1988) and Wolfe *et al* (1999), indirect standardization is usually preferred. Often, as with the data in this thesis, indirect standardization is expected to be the only option as stratum specific mortality rates are likely to be poorly estimated from the units of interest.

However, it is possible that some of the problems with direct standardization could be overcome. Ad-hoc rules could be used; for example, an early approach was to ignore those strata where rates are likely to be very poorly estimated (Ogle, 1886). An alternative, and more conservative, approach could be to derive estimates by assuming that no difference exists where rates cannot be estimated. However, this approach seems inferior, and possessing potential problems, when compared to using indirect standardization. Some more

of the practical differences between direct and indirect standardization will be discussed further in §5.5 when looking at the ratio of the observed and expected deaths. One further consideration is that, with the model used here, each unit has been compared to the rest of the Region. This means that each unit is compared to a different standard population, thus eliminating the advantage of direct standardization.

Various methods of comparing the total number of observed deaths $\sum_{i=1}^{n} d_i$ to the expected total $\sum_{i=1}^{n} \hat{\pi}_i$ are explored in the following Sections. These methods fall broadly into two types: significance tests and effect estimation. In the next Section significance testing approaches will be illustrated and discussed, with methods used to estimate effect sizes shown in later Sections.

5.4.2 Significance Tests with Standardized Outcomes

A naïve approach to reporting the difference between the observed and expected mortality of a unit would be to present the p-value for the observed number of deaths, under the null hypothesis that a unit has the same underlying mortality rates as the rest of the Region as quantified by the expected number of deaths. Such an approach has been used by the California Office of Statewide Health Planning and Development to categorise the performance of hospitals in the treatment of acute myocardial infarction (Healthcare Quality and Analysis Division, 2002), although this was only one part of a much larger, and robust, reporting procedure. Here, these methods will be illustrated using indirect standardization, because the expected number of deaths can be more reliably estimated. However, the methods outlined can equally be applied to estimates derived using direct standardization.

It is apparent from Table 5.7 that the inclusion of gestational age in the model produces values for the expected number of deaths that are generally closer to the observed values than those obtained without adjustment. However, there still remain some units that have a 'large' difference; for example Units 6 and 12. It may be of interest to investigate the statistical significance of these differences.

Although a method for calculating exact confidence intervals, and exact p-values, for observations from a binomial distribution was discussed in Chapter 2, that method cannot be used when observations have different event probabilities. Different predicted probabilities can occur after risk-adjusting for morbidity. It can easily be seen that an infant born at a very

early gestational age, say 22 weeks, is likely to have a much higher predicted probability of death that another born at 32 weeks, even if they are admitted to the same NICU. Therefore, infants admitted to any NICU are likely to have a range of predicted probabilities for death and any method to calculate the probability for the observed number of deaths, under the null hypothesis, will need to take this into account. Methods for estimating probabilities in such circumstances are now discussed.

Exact method

Luft & Brown (1993) proposed an exact method based on the probability of survival q_i where:

$$q_i = 1 - \pi_i$$

The probability that there are no deaths among N admissions to a NICU is $\prod_{i=1}^{N} q_i$ and, hence, the probability of at least one death is $1 - \prod_{i=1}^{N} q_i$.

The probability of exactly one death occurring is estimated by summing all of the possible combinations that one death could have occured in the observed data:

$$P(D=1) = \pi_1 . \prod_{i=2}^{N} q_i + q_1 . \pi_2 . \prod_{i=3}^{N} q_i + \prod_{j=1}^{2} q_j . \pi_3 . \prod_{i=4}^{N} q_i + \ldots + q_N . \pi_{N-1} . \prod_{j=1}^{N-2} q_j + \prod_{i=1}^{N-1} q_i . \pi_N$$

The probability of at least two deaths occurring is give by one minus the probability of zero deaths plus all the possible combinations one death could have occured:

$$P(D \geq 2) = 1 - \left(\prod_{i=1}^{N} q_i + \left[\pi_1 \prod_{i=2}^{N} q_i + \pi_2 q_1 \prod_{i=3}^{N} q_i + \ldots + \pi_N \prod_{i=1}^{N-1} q_i \right] \right)$$

This method can then be continued to find $P\left(D \geq \sum d_i\right)$.

The California Hospital Outcomes Project (Healthcare Quality and Analysis Division, 2002) calculated such a probability for each of the 398 hospitals investigated. However, when the observed number of deaths was less than the expected, they calculated the probability of the observed value or less. The probability of a greater or equal number of deaths was calculated for those hospitals with more observed deaths than expected. A hospital with an associated p-value of less than 0.01 was classified as being *"significantly better [worse] than expected"*.

The most obvious approaches to calculating such probabilities are difficult due to computing time and memory requirements. The number of calculations is dependent on both the number

of admissions and the number of deaths, and computing requirements grow quickly with increasing sample size. For each number of deaths (r), the number of ways of selecting r admissions from a total of n (combinations) is given by:

$$ {}_nC_r = \frac{n!}{(n-r)!\,r!} $$

For example, Unit 9, the smallest observed number of admissions (35) and deaths (1), only requires one calculation to calculate the probability of all admissions surviving to discharge and 35 to calculate the probability of exactly one death. However, once the sample grows to that of Unit 10 (146 admissions and 5 deaths) the number of individual probabilities to be summed becomes:

0 deaths:	1
1 death:	146
2 deaths:	10,585
3 deaths:	508,080
4 deaths:	18,163,860
5 deaths:	515,853,624

This is a total of 534,536,296 calculations.

For the small units such probabilities are straightforward to calculate in a statistical package such as SAS. `PROC PLAN` can be used within SAS to generate a dataset containing all combinations of a given set of observations for a specified size of sample. This can then be used to calculate the probability of observing the specified exact number of observation. However, for more than a small number of deaths, such a dataset with all possible combinations of observations of a given size is very large. Using Unit 10 as an example again, a data set containing all possible combinations of size five from 146 observations will have (515,853,624 × 5 =) 2,759,268,120 observations. The SAS program used for analysing data for this thesis has insufficient memory to generate such a dataset. An alternative programming approach would be to use 'loops' to generate one combination at a time, but such an approach can be expected to take a very long time to run.

Luft & Brown supplied an alternative, efficient approach to the calculations (Appendix D.3). Their method works by using cumulative sums of probabilities for different outcomes and full

details are given in their paper. They recommended that exact p-values be calculated if the number of observed deaths is 15 or less. Other authors have suggested that when the predicted number of deaths is equal to or exceeds five, approximation methods are adequate (Fleiss *et al,* 2003:26). Such methods will be discussed next and in the example below, using the TNS data, both exact and approximate p-values will be calculated for all units.

Normal Approximation

The probability that the i-th infant will die before discharge is π_i and its variance is given by $\pi_i(1-\pi_i)$. The true value of π_i is unknown but can be estimated by $\hat{\pi}_i$ and its variance by $\hat{\pi}_i(1-\hat{\pi}_i)$. If the observations within Unit *j* are assumed to be observations from a set of independent Bernoulli trials then $\sum_{i=1}^{n}\pi_i$ is estimated by $\sum_{i=1}^{n}\hat{\pi}_i$: its variance can be estimated by $\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)$. When $\sum\hat{\pi}$ and $\sum(1-\hat{\pi})$ are is sufficiently large (≥ 5), under the null hypothesis that $\Sigma\pi_i = \Sigma d_i$, the following distribution approximately holds (Luft and Brown, 1993; Fleiss *et al,* 2003:26):

$$\sum d_i \sim Normal\left(\sum \pi_i, \sum \pi_i[1-\pi_i]\right)$$

Thus:
$$z = \frac{\sum d_i - \sum \hat{\pi}_i}{\sqrt{\sum \hat{\pi}_i(1-\hat{\pi}_i)}}$$

As this approximation is using a continuous probability distribution to approximate a discrete distribution, the absolute difference $\left|\sum d_i - \sum \hat{\pi}_i\right|$ is often reduced by a continuity correction of the value ½ (Armitage and Berry, 1994; Bland, 1995:221; Fleiss *et al,* 2003:27). The calculated z-score can then be converted to a one-tailed probability using the standard normal distribution.

Comparison of exact and Normal approximation methods

For the TNS data, the calculated probabilities using the exact method proposed by Luft & Brown and using the Normal approximation (with the continuity correction) are shown in Table 5.8.

Table 5.8 P-values for observed mortality

Unit	Died		$\Sigma d_i - \Sigma \pi_i$	P-values[g]	
	Observed (Σd_i)	Expected ($\Sigma \pi_i$)		Exact method	Normal approx.
1	21	17.8	3.2	0.2176	0.2208
2	30	27.9	2.1	0.3181	0.3232
3	2	1.1	0.9	0.2717	0.3194
4	6	5.3	0.7	0.4457	0.4684
5	41	31.1	9.9	0.0277	0.0250
6	54	35.1	18.9	0.0001	0.0001
7	29	19.7	9.3	0.0104	0.0084
8	8	9.8	-1.8	0.3217	0.3132
9	1	0.9	0.1	0.6222	0.6768
10	5	8.9	-3.9	0.0892	0.0956
11	62	46.9	15.1	0.0051	0.0042
12	5	11.5	-6.5	0.0158	0.0227
13	3	6.0	-3.0	0.1159	0.1225
14	2	5.2	-3.2	0.0770	0.0873
15	10	9.9	0.1	0.5431	0.5566
16	6	4.9	1.1	0.3607	0.3804

For Units 9 & 15 the p-values are greater than 0.5 and, due to the definition of these one-sided p-values, this should not be expected. However, in the case of these two units the continuity correction used (½) is greater that the observed difference. In this case the application of the correction changes the sign of the difference, clearly an undesirable effect. However, by its nature this problem only occurs when the observed mortality and expected mortality are very close (i.e. an absolute difference of 0.5 or less) and there is no danger of misclassifying a unit as an outlier. One solution is to apply the correction only if the absolute value of the observed difference exceeds the value of the correction; i.e. 0.5 (Fleiss *et al,* 2003:27).

To illustrate the effect of the addition of the continuity correction, p-values calculated without it are shown in Table 5.9.

[g] $P(\Sigma d_i \leq \Sigma \pi_i)$ where $\Sigma d_i < \Sigma \pi_i$ & $P(\Sigma d_i \geq \Sigma \pi_i)$ where $\Sigma d_i > \Sigma \pi_i$

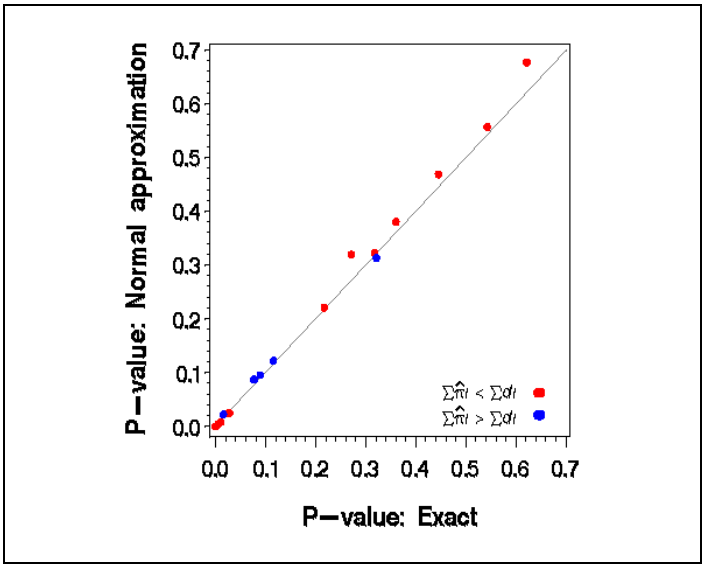
Table 5.9 Corrected p-values

Unit	$\Sigma d_i - \Sigma \pi_i$	P-values[h]		
		Exact method	Normal approx.	
			Continuity correction	No continuity correction
1	3.2	0.2176	0.2208	0.1794
2	2.1	0.3181	0.3232	0.2817
3	0.9	0.2717	0.3194	0.1418
4	0.7	0.4457	0.4684	0.3757
5	9.9	0.0277	0.0250	0.0194
6	18.9	0.0001	0.0001	0.0001
7	9.3	0.0104	0.0084	0.0057
8	-1.8	0.3217	0.3132	0.2473
9	0.1	0.6222	0.6768	0.4604
10	-3.9	0.0892	0.0956	0.0668
11	15.1	0.0051	0.0042	0.0032
12	-6.5	0.0158	0.0227	0.0150
13	-3.0	0.1159	0.1225	0.0815
14	-3.2	0.0770	0.0873	0.0542
15	0.1	0.5431	0.5566	0.4811
16	1.1	0.3607	0.3804	0.2891

There are several points to note from Table 5.9. First, as expected, all of the p-values calculated using the Normal approximation without the continuity correction are less than 0.5. Second, all of the p-values calculated using the Normal approximation with the continuity correction are greater than those calculated without the continuity correction, since the absolute value of the difference between observed and expected mortality has been reduced. In almost all cases, the p-values calculated using the Normal approximation with the continuity correction are closer to the exact p-values than those calculated without using the continuity correction.

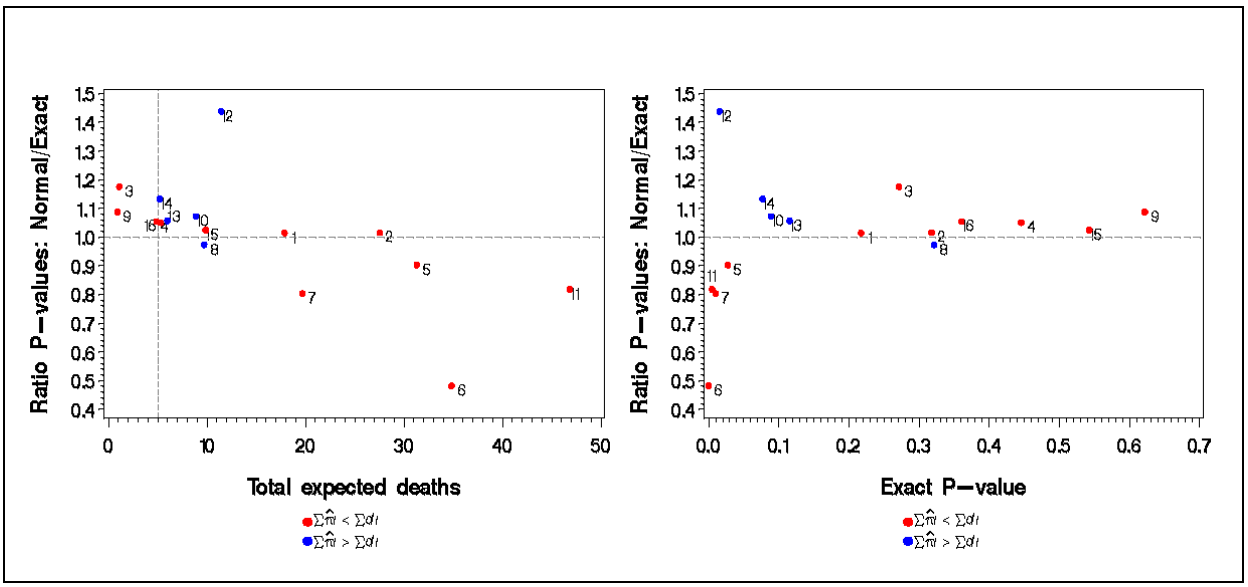
[h] $P(\Sigma d_i \leq \Sigma \pi_i)$ where $\Sigma d_i < \Sigma \pi_i$ & $P(\Sigma d_i \geq \Sigma \pi_i)$ where $\Sigma d_i > \Sigma \pi_i$

Figure 5.4 Plot of Normal approximation p-values against exact p-values



The absolute differences between the approximate p-values (with the continuity correction) and the exact values are small (Figure 5.4). However, there are substantial relative differences between the probabilities calculated from the two methods (Figure 5.5).

Figure 5.5 Ratio of p-values from Normal and Exact methods by total expected deaths



It is also of interest to note that these differences occur with units that have total expected deaths greater than five. It is unclear why this should be the case, but it may be due to the wide range in the value of π or, perhaps, because the p-value is small and a small absolute difference can result in a large relative difference. The problem arises in units with small p-values, those with the more extreme performance, and since these are usually the ones of interest and it is important that they are estimated correctly.

Further investigation would be required to try to find the cause of these differences. However, since the exact probabilities are straightforward to calculate, it is recommended that exact p-values be reported for all units where the total number of expected deaths is less than 20.

Other methods

Although the exact method outlined above is an appropriate method to estimate such p-values other methods have been suggested. These are briefly described next.

Lexian Distributions

The family of Lexian distributions specify the distribution of a set of binary outcomes where the event probabilities are not necessarily equal but, rather, follow a probability distribution themselves. The probability density function for the Lexian distributions is:

$$P(O=o)=\int_0^1 \binom{n}{o}\theta^o(1-\theta)^{n-o} f(\theta)d\theta \qquad (5.17)$$

where (Stuart and Ord, 1994:172):

$$E(O)=nE(\theta)$$

$$Var(O)=nE(\theta)[1-E(\theta)]+n.(n-1)Var(\theta)$$

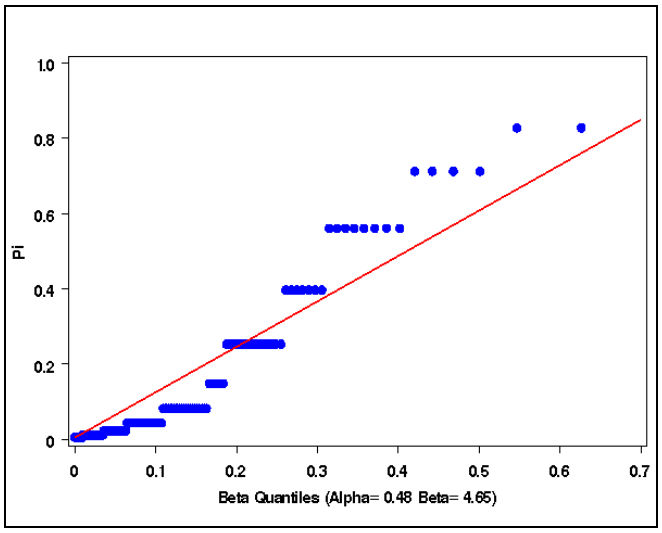
$$=n\bar{\pi}.(1-\bar{\pi})+n(n-1)Var(\theta)$$

It can be seen that in the special case where θ is fixed, i.e. $Var(\theta)$=0, the above expressions reduce to those of the Binomial distribution. It can also be seen that the variance is greater by the value of $n(n-1)Var(\theta)$ than that of the Binomial distribution with the same mean.

One consideration with this distribution is the choice of probability distribution for θ. Both the Beta and Poisson distributions have been suggested (Edwards, 1960). In previous work, Edwards had assumed a Beta distribution for the event probabilities (Edwards, 1958). He felt that since the variance of π in his data was small, mispecification of this distribution would have a minor influence on the results. This is not the case with the TNS data used in this thesis. The range of estimated probabilities within each NICU is large; for example in Unit 1 $\hat{\pi}$ ranges from 0.0063 to 0.8280. The estimated probabilities seem a poor fit to the beta distribution; for example the Beta Q-Q plot for Unit 1 (using maximum likelihood estimates for α and β) is shown in Figure 5.6. Inspection of this plot suggests that the observed

distribution has longer tails that that expected from the corresponding Beta distribution. All the other NICUs show a similar pattern.

Figure 5.6 Q-Q Plot for estimated π for Unit 1



Obviously, the distribution of $\hat{\pi}$s will depend on the model used to estimate them. In this example, they follow the same distribution as the observed gestational ages, since this is the only variable in the model. It is possible that the addition of other variables to the model may produce an observed distribution close to that of a Beta distribution. However, this is not necessarily certain in all cases.

The difficulty of specifying a distribution for the estimated probabilities means that this approach will not be considered further in this thesis.

Poisson Approximation

When the events are rare it may be assumed that the number of observed deaths Σd_i follows a Poisson Distribution (Luft and Brown, 1993):

$$P(D \geq \Sigma d_i) = 1 - \left(e^{-\Sigma\pi_i} + e^{-\Sigma\pi_i}.\Sigma\pi_i + \frac{e^{-\Sigma\pi_i}.[\Sigma\pi_i]^2}{2!} + \ldots + \frac{e^{-\Sigma\pi_i}.[\Sigma\pi_i]^{\{\Sigma d_i\}-1}}{[\{\Sigma d_i\}-1]!} \right)$$

This follows from the recognition that $\sum_{i=1}^{n} \pi_i = \lambda$ and that if π_i is small then $\sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i)$ approaches $\sum_{i=1}^{n} \hat{\pi}_i$, hence both the expectation and variance of the sum of events take the value λ.

However, as well as the assumption that $\Sigma\pi_j \geq 5$, this approximation only holds if all the π_j are small and relatively uniform (Luft and Brown, 1993). It is clear that this is not the case for the TNS data used in this thesis. Therefore, this method will not be considered further.

Simulation

One further method suggested involves drawing, for each observation, a value from the Uniform distribution, *Uniform*(0,1). If this value is greater than the predicted probability of dying for that infant, then the observation is counted as a 'death'. Once this has been repeated for all observations from the unit of interest, the total number of simulated 'deaths' is recorded. This process is repeated many times and the proportion of times that the total number of simulated 'deaths' equals or exceeds the observed value is the estimated probability of observing that many or more deaths (Luft and Brown, 1993).

Although this method provides an unbiased estimate for the required p-value, its precision depends on the number of simulations. This can become computationally intensive and slower to run than the exact method outline above. Luft & Brown compared the two methods using data from 34,234 patients from 465 hospitals and calculated exact values for all hospitals with 15 or fewer observed deaths. They found that the exact method took less than 90 seconds whereas the simulation method took around 2¼ hours, using a Macintosh II with a 68881 floating point coprocessor and 5Mb of memory (Luft and Brown, 1993). Although computing resources have developed greatly since their study, it is difficult to see any advantage in using the simulation method when the exact method is so straightforward. It will, therefore, not be considered further here.

Chi-square Test

Knaus *et al* (1993) used the chi-squared test (with one degree of freedom) to determine the statistical significance of the difference between the observed and expected mortality rates for each unit. This is illustrated using the TNS data in Table 5.3.

Three units (Units 6, 7 and 11) have p-values of less than 0.05, the cut-off used by Knaus, and all of these have expected mortality totals greater than the observed. It can also be seen that the p-values shown above are similar to the p-values for the odd ratios shown in Table 5.3.

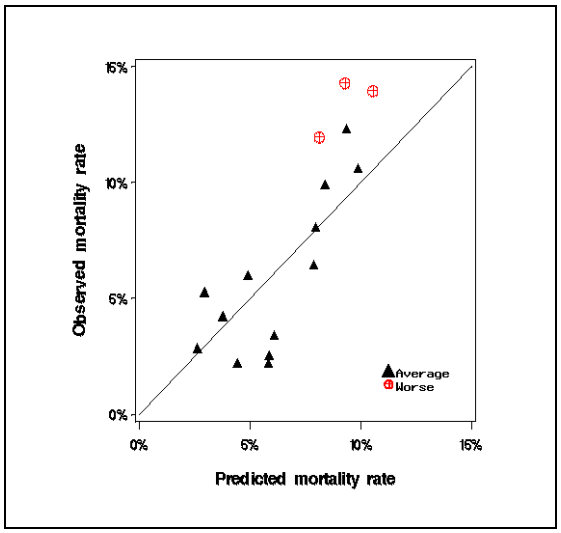
One further point to note is that the p-values calculated using the chi-squared distribution are, by definition, two-sided p-values. The p-values reported by the California Office of Statewide Health Planning and Development mentioned at the start of §5.4.2 are one-side, as were the others described above.

Table 5.10 Chi-square p-values

Unit	Died			p-value
	Observed (Σd_i)	Expected ($\Sigma \pi_i$)	$\frac{(\sum d_i - \sum \hat{\pi}_i)^2}{\sum \hat{\pi}_i}$	
1	21	17.8	0.594	0.44
2	30	27.9	0.151	0.70
3	2	1.1	0.707	0.40
4	6	5.3	0.080	0.78
5	41	31.1	3.129	0.077
6	54	35.1	10.203	0.0014
7	29	19.7	4.344	0.037
8	8	9.8	0.314	0.57
9	1	0.9	0.008	0.93
10	5	8.9	1.701	0.19
11	62	46.9	4.844	0.028
12	5	11.5	3.637	0.056
13	3	6.0	1.502	0.22
14	2	5.2	2.004	0.16
15	10	9.9	0.002	0.96
16	6	4.9	0.251	0.62

As a point of interest, the authors then used these p-values to label points in a scatterplot of observed mortality rates against expected mortality rates. This is illustrated using the TNS data in Figure 5.7.

Figure 5.7 Plot of observed against expected mortality rate



The units performing worse than expected are those with high rates of predicted mortality. This may indicate a lower quality of care, but may also be the result of inadequate adjustment for infant morbidity at admission. In fact, while the truth about the former is unknown, we know that the latter is true; gestational age alone is unlikely to be adequate case-mix adjustment.

While such an approach may be useful in an initial inspection of the data, the lack of estimated confidence intervals means that it is an insufficient approach on its own.

To *"... determine the amount of variation across ICUs that was accounted for by predictions ..."* the authors also used a least-squares linear regression model, with the observed mortality rate as the outcome and the expected rate as the covariate, to estimate the coefficient of determination (R^2). When using the TNS data $R^2 = 0.70$. Thus some 70% of the variation in observed mortality rates is 'explained' by differences in gestational age between the units.

5.5 Standardized Outcome Ratios

In the previous Section statistical methods were described and illustrated to carry out significance tests on standardized outcomes. No matter which of these methods is used, a p-value only gives information about the statistical significance of a difference between the observed and expected. What is of more interest is an estimate of the clinical size of this difference, and the uncertainty around it. Such estimates can be obtained by comparing the expected number of deaths to the observed number in different ways. Two of the most used approaches will be discussed in the rest of this Chapter. First, the ratio of the observed and expected deaths is investigated and then their difference is briefly considered.

The ratio between the observed number of deaths and the expected number (obtained through standardization) can be used as a summary measure of a unit's outcome. The two most popular approaches, the **comparative mortality figure** and the **standardized mortality ratio**, are described next.

5.5.1 Comparative Mortality Figure

The comparative mortality figure (CMF) is derived using direct standardization and is the ratio of the expected number of deaths in the reference population to the observed number (Fleiss *et al,* 2003:639):

$$CMF = \frac{Expected_R}{Observed_R} = \frac{\sum_{i=1}^{l} \pi_i p_{Ri}}{\sum_{i=1}^{l} \pi_{iR} p_{Ri}} \qquad (5.18)$$

where: π_i is the probability of death for an observation in stratum i in unit of interest

π_{Ri} is the probability of death for an observation in stratum i in reference population

p_{Ri} is the proportion of observations in stratum i in reference population

The value unity represents no difference between the observed and expected totals. A value greater than unity indicates that the expected number is greater than the observed, so that the stratum specific mortality rates in the unit of interest are, in some way, greater than those of the reference population. A CMF value of less than unity represents the opposite scenario.

This ratio may be the most appropriate summary statistic to use as the weights (p_{Ri}) are the same for each unit and, therefore, allow direct comparisons between different units to be made in addition to the comparison of a unit to the reference population. This is made clearer by supposing that mortality rates are to be compared between two populations a and b by reference to a standard population. Silcock (1959) suggested three properties (the first two essential and the third desirable) for such a summary statistic if it is to be used to compare the outcome between the two health care providers:

where: π_{xi} is the mortality rate for stratum i in population x;

p_{xi} is the proportion of population x in stratum i;

Π_x is the overall death rate in population x (e.g. $\Pi_a = \sum_{i=1}^{n} \pi_{ai} p_{ai}$).

Property 1

If $\alpha \le \frac{\pi_{ai}}{\pi_{bi}} \le \beta$

then $\alpha \le \frac{\Pi_a}{\Pi_b} \le \beta$

This property specifies that the value of the ratio of the overall mortality rates between two populations lies within the limits of the ratios of the stratum specific rates.

Property 2

If $\Pi_a \neq \Pi_b$ then the inequality should be due to $\pi_{ai} \neq \pi_{bi}$ for some or all *i*, and to nothing else.

Property 3

"... the comparative function $[\Pi_a / \Pi_b]$ *should have a meaning other than in the abstract mathematical sense in which a number is 'explained' by pointing to the mathematical formula from which it was derived."* (Silcock, 1959) That is to say, there should be a clear clinical interpretation to such a statistic. An example of a summary statistic fulfilling this property (CMF) and one not (SMR) are shown below.

The comparative mortality figure fulfils all three of these properties.

That *Property 1* holds can be seen by considering that if $\alpha \leq \frac{\pi_{ai}}{\pi_{bi}} \leq \beta$ then:

$$\alpha p_{Ri}\pi_{bi} \leq p_{Ri}\pi_{ai} \leq \beta p_{Ri}\pi_{bi}$$

hence:
$$\alpha = \frac{\alpha\sum_{i=1}^{n} p_{Ri}\pi_{bi}}{\sum_{i=1}^{n} p_{Ri}\pi_{bi}} \leq \frac{\sum_{i=1}^{n} p_{Ri}\pi_{ai}}{\sum_{i=1}^{n} p_{Ri}\pi_{bi}} \leq \frac{\beta\sum_{i=1}^{n} p_{Ri}\pi_{bi}}{\sum_{i=1}^{n} p_{Ri}\pi_{bi}} = \beta$$

and
$$\alpha \leq \frac{\Pi_{ai}}{\Pi_{bi}} \leq \beta \text{ as required.}$$

If $CMF_a \neq CMF_b$, then from (*5.7*), $\sum_{i=1}^{n} p_{Ri}\pi_{ai} \neq \sum_{i=1}^{n} p_{Ri}\pi_{bi}$. Since p_{Ri} is common to both sides this inequality is due to $\pi_{ai} \neq \pi_{bi}$ for one or more *i*, thus satisfying the condition for *Property 2*.

The third property can be seen to be met by considering $CMF_a / CMF_b = 1 + h$. In this case h can be interpreted by seeing that if the reference population had the same mortality rates as population *a* then there would have been 100*h*% more deaths than if it had the mortality rates of population *b*.

However, the problem remains of sufficiently precise estimates of strata specific death rates being able to be estimated and these properties also depend on the same reference population

being used for all comparisons. Neither of these conditions holds for the analyses in this thesis.

5.5.2 Standardized Mortality Ratio

An alternative summary statistic, estimated using indirect standardization, is the standardized mortality ratio (SMR):

$$SMR = \frac{Observed}{Expected} = \frac{\sum_{i=1}^{I} \pi_i p_i}{\sum_{i=1}^{I} \pi_{Ri} p_i} \qquad (5.19)$$

where: π_i is the probability of death for an observation in stratum i in unit of interest

p_i is the proportion of observations in stratum i in unit of interest

p_{Ri} is the proportion of observations in stratum i in reference population

This is the ratio of the observed to the expected number of deaths for each unit. Although each SMR is a true measure of the difference between each population and the reference population, given the population structures, it can be seen that the SMR for each unit is weighted according to its own population structure, i.e. $p_1, p_2, \ldots, p_I$. This means that it may not be possible to compare the SMRs of two units even if they are standardized to the same reference population. The three properties proposed by Silcock to allow the comparison of mortality ratios were discussed in §5.5.1, and these will now be considered in relation to the standardized mortality ratio.

Algebraic details are shown in Appendix C.2 that *Property 1* (that the overall ratio between two populations of interest does not take a value more extreme than any of the stratum specific ratios) does not hold for the SMR, but perhaps it can more easily be shown by a counter-example. Consider two populations, X and Y, whose death rates are to be indirectly standardized by sex to a reference population (Table 5.11).

Table 5.11 Internal comparison using SMRs

Population	**Ref.**	**X**			**Y**		
	Rate	**n**	**obs**	**exp**	**n**	**obs**	**exp**
Male	0.2	700	350	140	300	180	60
Female	0.4	300	90	120	700	240	280
Total			**440**	**260**		**420**	**340**

The stratum specific SMRs are:

Male: $SMR_{X.male} = 2.50$ $SMR_{Y.male} = 3.00$ $\frac{SMR_{X.male}}{SMR_{Y.male}} = 0.83$

Female: $SMR_{X.female} = 0.5$ $SMR_{Y.female} = 0.86$ $\frac{SMR_{X.female}}{SMR_{Y.female}} = 0.88$

Overall: $SMR_X = 1.69$ $SMR_Y = 1.24$ $\frac{SMR_X}{SMR_Y} = 1.37$

Therefore, in this example the ratio of the overall SMRs is greater than the ratio of either stratum specific SMRs. Moreover, it can be see that Population X has lower stratum specific death rates that Population Y ($\pi_{X.male} = 0.50; \pi_{Y.male} = 0.60;\ \pi_{X.female} = 0.30;\ \pi_{Y.female} = 0.34$), but that the overall death rate is higher in X ($\Pi_X = 0.44; \Pi_Y = 0.42$). This is a version of 'Simpson's Paradox' (Simpson, 1951; Heydtmann, 2002). Not only is the overall ratio of rates more extreme than either of the stratum specific rates but, in this case, the conclusion that Y has a higher overall death rate is not supported by either of the stratum specific rates. This is clearly an undesirable characteristic and shows that the SMR does not fulfil *Property 1*.

Property 2 states that any inequality between two SMRs is solely due to differences in the stratum specific death rates. However, we have:

if $SMR_a \neq SMR_b$

then $$\frac{\sum_{i=1}^{l} \pi_{ai} p_{ai}}{\sum_{i=1}^{l} \pi_{Ri} p_{bi}} \neq \frac{\sum_{i=1}^{l} \pi_{bi} p_{bi}}{\sum_{i=1}^{l} \pi_{Ri} p_{bi}}$$

From this it can only be concluded that either $\pi_{ai} \neq \pi_{bi}$ or $p_{ai} \neq p_{bi}$, or that both are true for at least one *i*. The weights used in the denominator of the SMR are specific to the unit being investigated and, therefore, the ratio estimated for each provider is weighted (biased) in relation to the characteristics of their populations (Bhopal, 2002:194-198). Since the observations in each unit are likely to have different empirical distributions for the variables used in standardization (risk-adjustment), the rates estimated for each unit will, strictly, not be comparable with each other. The only comparison that truly can be made is the comparison between the population under study and the reference population.

The final property, *Property 3*, requires that a clinical interpretation can be put to the comparison of two SMRs. This is clearly not possible for the reasons outlined above; that is,

the difference may be due to either differences in stratum-specific mortality rates or because of different population structures.

Although none of the three properties advocated are met when comparing SMRs between two populations of interest, there is no problem when comparing a population of interest to the reference population. While the use of the SMR has a long history (Neison, 1844; Keiding, 1985), the size of the errors produced when comparing SMRs to each other is unclear (Howell, 2002). The size of the errors is a function of the differences in both the population structure and the stratum specific rates, and this will be investigated in the next Section. However, what is clear, is that caution is required when comparing units according to the magnitude of their SMRs (Howell, 1995).

Standardized mortality ratios are sometimes presented in a number of different ways. In some cases the SMR is multiplied by 100 and expressed as a percentage (Armitage and Berry, 1994:439). An alternative method of presentation that has been published is to use a percentage difference from the expected (Zullini *et al,* 1997):

$$\textbf{\textit{Percentage difference}} = \frac{(Observed - Expected)}{Expected} \times 100$$

A further adaptation, used by the New York Department of Health, is to multiply the mortality ratio by a measure of the overall mortality rate in order to provide a risk-adjusted rate for each provider (New York State Department of Health, 1998a). However, since all of these are simply a rescaling of the SMR, and do not add any further information, they will not be considered further in this thesis.

5.5.3 Comparison of the SMR and CMF

Indirect standardization has been proposed for the data in this thesis, since the use of direct standardization is not viable once additional risk-adjustment variables are added to the model. As discussed above, it is argued that only comparisons between each neonatal unit and the reference population can be made using the SMR. It has been suggested that this is the only comparison of interest, as people are not interested in direct comparisons between hospitals:

> *"Most patients are probably not interested in where exactly their hospital is in the league table, but they are interested, and rightly so, in knowing that their hospital constantly monitors its performance and acts immediately if there is evidence that it is*

> *not doing well."* (Society of Cardiothoracic Surgeons of Great Britain and Ireland, 2004)

While these are good sentiments, they may underestimate people's curiosity. Despite the known limitations of the SMR, it is enormously tempting to compare SMRs between units when several are published together. This thesis will present the SMRs for all 16 units in a single table. It is felt to be important that the figures shown are not grossly misleading, even if used in an inappropriate manner. For this reason the similarity or otherwise between the two statistics will be discussed in this section. Notwithstanding the use of different reference populations in this thesis for each unit (i.e. the other 15 units), the problem with the SMR is that the weights used to calculate it are different for each unit (*5.19*). To some, such characteristics immediately mean that such ratios are unsuitable:

> *"Any so-called method of standardization which does not fulfil this condition* [the ratios being comparable with each other] *hardly deserves the name at all: it is only a 'single-pair' method, and if it is applied to a number of groups it may only be thanks to the mercy of Providence that it is not grossly misleading."* (Yule, 1934)

M statistic

However, the size of the bias introduced is unknown. A statistic, ***M***, has been proposed to try to quantify the difference in case-mix between two populations that may be useful in comparing two populations: for example *a* & *b* (Hollis *et al,* 1995). If both populations are divided into *K* intervals according to the value of the predicted probability of death π_i, the *M* statistic is given by:

$$M = \sum_{k=1}^{K} \boldsymbol{minimum}\left(\boldsymbol{F_k}, f_k\right) \qquad (5.20)$$

where: F_k proportion of observations in population *a* in interval *k*

f_k proportion of observations in population *b* in interval *k*

The intervals suggested are shown in Table 5.12. Using such intervals, it has been proposed that a value of less than 0.88 indicates that the two populations are 'significantly' different, although no justification was given for such a choice (Boyd *et al,* 1987). However, such a statistic is only able to quantify the difference rather than to indicate whether the difference will affect the results (Hollis *et al,* 1995).

Table 5.12 Intervals for M statistic

Interval	$(1-\pi_i)$ range
1	0.96 – 1.00
2	0.91 – 0.95
3	0.76 – 0.90
4	0.51 – 0.75
5	0.26 – 0.50
6	0.00 – 0.25

Weighted standardized mortality ratios

Various methods of weighting the SMR have been proposed in order to try to overcome the problem of non-comparability. One approach is derived from the W_S statistic (Hollis *et al,* 1995), also sometimes called the standardized Z score (Tibby *et al,* 2002). Although Hollis *et al* illustrated their method using the difference between the observed and expected mortality, it can equally be applied to the ratio, in effect producing a *"standardized SMR"* (Glance *et al,* 2000). In order to obtain this variation of the SMR, the observations from the reference population are categorised into K intervals according the values of the predicted probability of death. The proportion of reference observations in each interval j is F_j, hence $\sum_{j=1}^{K} F_j = 1$. Interval specific SMRs are then calculated (SMR_j) and weighted using $F_1, F_2, \ldots, F_K$:

$$\textit{Standardized SMR} = \sum_{j=1}^{K}\left(\frac{\sum_{i=1}^{n_j} d_i}{\sum_{i=1}^{n_j}\hat{\pi}_i}.F_j\right) = \sum_{j=1}^{K}\left(SMR_j.F_j\right) \qquad (5.21)$$

However, a decision is required on how many intervals and what cut-off values should be used. In their paper Hollis *et al* (1995) used the six intervals shown in Table 5.12. However, the choice of intervals may influence the estimated overall SMR. A further problem is that, for the smaller units, many of these intervals will have no observed deaths and, therefore, an estimated interval specific SMR of zero. Even in the intervals for which there are observed deaths, the SMRs are likely to be poorly estimated because of the small numbers.

This is illustrated with the TNS data, using the same model used to obtain the SMRs as in §5.5.2. The cut-off values $0.00 \leq \hat{\pi}_i \leq 0.15$, $0.15 < \hat{\pi}_i \leq 0.40$, $0.40 < \hat{\pi}_i \leq 0.60$, $0.60 < \hat{\pi}_i \leq 1.00$ were chosen to try to obtain roughly equal numbers of deaths in each interval. The estimates

were compared to the SMR (Table 5.13). For many units there are substantial differences in both the point estimates and the limits of the confidence intervals.

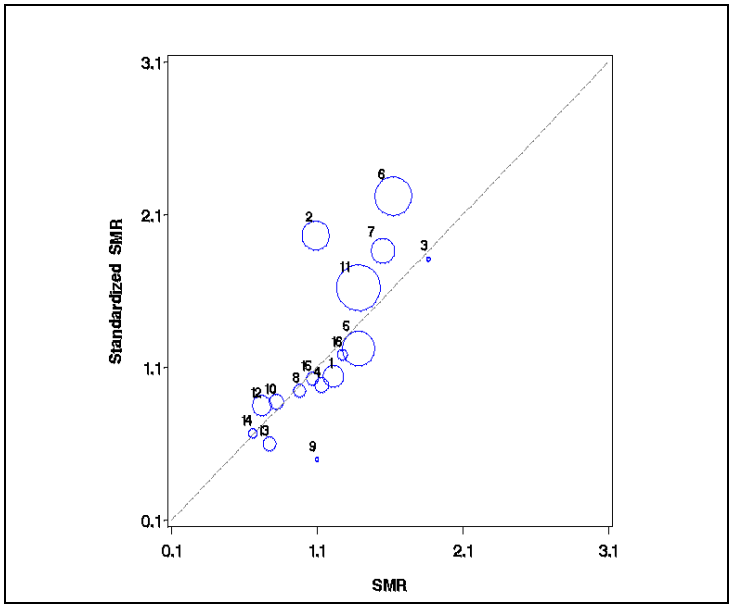
Table 5.13 SMR and standardized SMR by unit with 95% confidence intervals

Unit	SMR	(95% CI)	Standardized SMR	(95% CI)
1	1.18	(0.84 to 1.51)	0.95	(0.28 to 1.73)
2	1.07	(0.73 to 1.44)	1.92	(1.14 to 2.72)
3	1.80	(0.00 to 4.87)	1.74	(0.00 to 5.73)
4	1.12	(0.42 to 1.83)	0.83	(0.05 to 1.83)
5	1.32	(1.00 to 1.65)	1.19	(0.70 to 1.75)
6	1.54	(1.22 to 1.88)	2.20	(1.34 to 3.01)
7	1.47	(1.02 to 1.91)	1.80	(0.99 to 2.73)
8	0.82	(0.37 to 1.32)	0.76	(0.07 to 1.77)
9	1.09	(0.00 to 2.73)	0.26	(0.00 to 0.41)
10	0.56	(0.11 to 1.06)	0.64	(0.03 to 1.38)
11	1.32	(1.06 to 1.59)	1.54	(1.03 to 2.07)
12	0.44	(0.08 to 0.85)	0.60	(0.16 to 1.34)
13	0.50	(0.00 to 1.05)	0.33	(0.00 to 1.04)
14	0.38	(0.00 to 1.04)	0.39	(0.00 to 1.28)
15	1.01	(0.54 to 1.52)	0.92	(0.16 to 1.82)
16	1.23	(0.40 to 2.24)	1.16	(0.09 to 5.54)

The 95% confidence intervals were estimated using the percentile bootstrap method (introduced in §3.3.1). Methods to estimate confidence intervals for the SMR are explored further in §5.6.

The values for the SMR are also shown in Figure 5.8 where the size of the symbol for each unit is proportional to its number of admissions.

Figure 5.8 SMR and standardized SMR by unit size



It is not only the small units where the 'standardized SMR' differs substantially from the SMR: Units 2 and 6, for example, show large differences. The reason for this can be seen by looking at Unit 6:

Table 5.14 Components of standardized SMR for Unit 6

Interval (*j*)	$\sum \hat{\pi}_i$	$\sum d_i$	SMR_j	F_j	$SMR_j * F_j$
$0.00 \leq \hat{\pi}_i \leq 0.15$	7.90	19	2.40	82.4	1.98
$0.15 < \hat{\pi}_i \leq 0.40$	14.97	20	1.34	10.7	0.14
$0.40 < \hat{\pi}_i \leq 0.60$	5.42	5	0.92	4.1	0.04
$0.60 < \hat{\pi}_i \leq 1.00$	6.80	10	1.47	2.8	0.04

To ensure at least one observed death in each interval, the weights (F_j) given to the intervals vary greatly, and the first interval greatly influences the final estimated 'standardized SMR'. This results in the outcomes of infants with good prognoses having a large effect on the overall statistic, as noted previously (Younge *et al,* 1997). Whilst this is correct in the sense that this interval comprises most of the population, the interval only comprises 35% of the deaths observed in Unit 6. It may be possible to find an optimum set of cut-off values for the larger units, but this is likely to be impossible for the smaller ones. A further difficulty with the standardized SMR is that such a statistic has no simple interpretation. As has been discussed previously, it is helpful if a summary statistic has a simple interpretation. This approach may have a role in exploring differences in performance for different types of

infants but such sub-group analyses are likely to be more useful if the groups are derived using clinical criteria rather than the estimated predictive probabilities.

Other methods of weighting the SMR have been proposed. The Harmonically Weighted Ratio (HWR) for population *j* when comparing *J* populations to a standard population over *I* strata is given by (Lee, 2002):

$$HWR_j = \frac{\sum_{i=1}^{I}\left[\left(\sum_{k=1}^{J}\frac{1}{E_{ik}}\right)^{-1}\frac{O_{ij}}{E_{ij}}\right]}{\sum_{i=1}^{I}\left(\sum_{k=1}^{J}\frac{1}{E_{ik}}\right)^{-1}} \qquad (5.22)$$

where: O_{ij} is the observed number of events in strata *i* for population *j*

E_{ij} is the predicted number of events in strata *i* for population *j*

An alternative, also proposed by Lee, is the Geometrically Averaged Ratio (GAR) (Lee, 1999). For Unit *j*, using *K* risk strata, the Geometrically Averaged Ratio (GAR_j) is given by:

$$GAR_j = \exp\left(\frac{1}{K}\sum_{k=1}^{K}\log\frac{\sum_{i=1}^{n_k} d_i}{\sum_{i=1}^{n_k}\pi_i}\right) = \left(\prod_{k=1}^{K}\frac{\sum_{i=1}^{n_k} d_i}{\sum_{i=1}^{n_k}\pi_i}\right)^{\frac{1}{K}} \qquad (5.23)$$

However, all of these methods, while perhaps useful in some circumstances, attempt to solve a problem that may not be a significant problem in practice.

Differences between CMF and SMR

It has been noted that the values of the CMF and SMR are often similar (Breslow and Day, 1987:73). A function closely approximating the relationship between the CMF and SMR (for population *a*) has been proposed (Silcock, 1959):

$$CMF_a \approx SMR_a(1+\phi) \qquad (5.24)$$

where:

$$\phi = \sum_{i=1}^{n}(p_{Ri} - p_{ai})\left(\frac{\pi_{ai}}{\overline{\pi}_a} - \frac{\pi_{Ri}}{\overline{\pi}_R}\right) \qquad (5.25)$$

The condition where the CMF and SMR are approximately equal, i.e. when ϕ is small, can be found from (*5.25*). These conditions are:

- The differences in population structure are small, i.e. the values of $(p_{Ri} - p_{ai})$ are small;
- The stratum specific mortality rates are of similar magnitude across all strata; $\pi_{ai} \approx k\pi_{Ri}$ and $\overline{\pi}_{ai} \approx k\overline{\pi}_{Ri}$, hence $\left(\frac{\pi_{ai}}{\overline{\pi}_a} - \frac{\pi_{Ri}}{\overline{\pi}_R}\right)$ is small;
- Even if these two conditions do not hold, ϕ will still be small as long as they are not correlated, since $\sum_{i=1}^{n}(p_{Ri} - p_{ai}) = 0$ and $\sum_{i=1}^{n}\left(\frac{\pi_{ai}}{\overline{\pi}_a} - \frac{\pi_{Ri}}{\overline{\pi}_R}\right) = 0$.

Therefore, for the SMR to differ greatly from the CMF all three of the conditions must not hold: i.e. there must be differences in population structure, differences in the relative sizes of the stratum specific mortality rates and these differences must be strongly correlated. If all of these conditions do hold then the SMR is a good approximation to the CMF. This is a useful characteristic since it is legitimate to use the CMF to compare providers. Hence, although it is technically wrong to use the SMR, perhaps it can, with caution, be used in practice. Such an argument has been used in previous clinical papers, for example Takemura *et al* (1998).

Although it has been suggested that comparative mortality figures are likely to be poorly estimated from the TNS data it was still possible to obtain them from Table 5.6. These were then compared to the estimated standardized mortality ratios (Table 5.7). These values are shown in Table 5.15 and are plotted in Figure 5.9

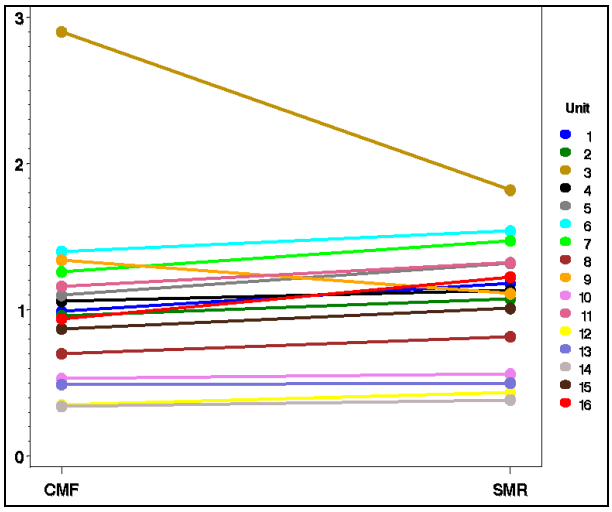
For most units there was close agreement between the estimates for the CMF and the SMR. There were only two units where the rank changed by more than one place: Unit 9 changed six places and Unit 16 changed four places. These two, together with Unit 3, showed the most difference between the two values. However, these are all small units and it seems reasonable to assume that the CMFs were poorly estimated; for example in Unit 9 there was a single death from 35 admissions. This comparison suggests that in this example, the differences between the CMF and the SMR were due to poor estimation of the CMF in very small units. However, gestational age was the only risk-adjustment variable considered. The addition of other variables into the model may introduce further differences in population structure between the units and, therefore, bias into the estimates for the SMRs. While this cannot be discounted in these data, the examples discussed earlier offer no evidence that this is likely to be a major problem.

Table 5.15 Comparative mortality figure and standardized mortality ratio

Unit	Comparative mortality figure	Standardized mortality ratio
1	0.99	1.18
2	0.96	1.07
3	2.90	1.80
4	1.06	1.12
5	1.10	1.32
6	1.40	1.54
7	1.26	1.47
8	0.70	0.82
9	1.34	1.09
10	0.53	0.56
11	1.16	1.32
12	0.35	0.44
13	0.49	0.50
14	0.34	0.38
15	0.87	1.01
16	0.94	1.23

The choice between CMF and SMR is a trade off between precision and bias (Breslow and Day, 1987:72). While there are those who believe that the gain in precision achieved from using the SMR as opposed to the CMF outweighs the inherent bias (Court and Cheng, 1995), there are those who take the opposite view (Julious *et al,* 2001; Rixom, 2002).

Figure 5.9 Comparative mortality figure and standardized mortality ratio



Use of the SMR

Even if the decision is taken that, in practice, the differences in the population structure of the units are unlikely to bias the results 'significantly', any comparison must come with a caution about the assumptions made. It is of concern that this caution is often missing, and almost always ignored. For example, the Dr Foster organization annually publishes indirectly standardized mortality ratios for UK hospitals trusts that appear in The Sunday Times newspaper (The Sunday Times, 2004a). Using these figures, high and low performing trusts were identified by the newspaper, for example:

> *"The worst overall was Royal Bournemouth and Christchurch Trust where 29% more patients died between July 2002 and July 2003 than would be expected after adjusting for their medical background and other factors."*

While it may be true that Royal Bournemouth and Christchurch Trust had the highest estimated SMR, it has been shown that this conclusion is dependent on the population structure of the Trust's patients, together with the population structures of the other Trusts. The error is further reinforced by the recommendation that:

> *"In this way, it is possible to produce a standardised figure that allows an assessment of the relative rates of mortality."* (The Sunday Times, 2004b)

To be fair to The Sunday Times this is an assertion made by the Dr Foster organisation itself:

> *"Standardisation of the ratio allows valid comparison between different hospitals serving different communities."* (Dr Foster, 2004)

In their publication of mortality following coronary artery bypass surgery, the New York State Department of Health report the SMR for individual surgeons and hospitals (multiplied by the overall mortality rate). This, they say, produces *"... the best estimate…of what the provider's mortality rate would have been if the provider had a mix of patients identical to the statewide mix."* (New York State Department of Health, 1998a). However, as has been shown, this is likely not to be the case.

It is also important to recognise that any predictive model used to risk-adjust is likely to perform better (more accurate calibration) for some patient groups than others (Glance *et al,* 2000). In particular, model coefficients for characteristics sparsely represented in the reference population may be particularly poorly estimated. It has been suggested that the patient characteristics in the index population should be compared to the population in which

the model was derived using statistical tests, and the model not applied if there is statistical evidence of a difference (Jones *et al,* 1995). However, while a difference may also lead to the problems outlined above, this too is not thought to represent a major problem in practical situations (Glance and Osler, 2001).

Use of the SMR in this thesis

The SMR was preferred over the CMF in this thesis. The comparison of each unit to the rest of the Region was of primary importance, rather than directly comparing the individual units. In any case, it has been shown in this Section that if such comparisons are made the bias introduced by the different population structures in the units is unlikely to result in any comparison of SMRs being grossly misleading.

One final problem is that there is no single, universally accepted, method for the estimation of confidence intervals for the standardized mortality ratio. Various suggested methods will be discussed and illustrated in the next Section.

5.6 Confidence interval for Standardized Mortality Ratio

There is no standard method for estimating a confidence interval for a standardized mortality ratio. Several methods have been suggested and these will be reviewed in this Section, along with the development of a Bayesian alternative.

The most commonly used estimation method is an approximation to the Normal distribution, but this approach assumes that the expected number of deaths is known exactly, and that all uncertainty arises from the observed deaths. This is unlikely to be strictly true. The expected number of deaths is usually derived from the parameter estimates of a logistic regression model. Since the data used within this model (in this case the rest of the Region) can be seen as a sample from a larger population of possible admissions, there is sampling variation associated with these parameter estimates and, therefore, with the expected number of deaths. While this uncertainty is likely to be small relative to that associated with the observed number of deaths, because of the relative sizes of the two data sets, its omission nevertheless may lead to inappropriately narrow confidence intervals. Two extensions to the Normal approximation method have been suggested to take the whole uncertainty into account and these will be discussed below. However these three methods are all based on the Normal approximation and, as discussed in §5.4.2, these may be poor estimates, particularly with

small samples. Three further methods will also be investigated that are not based on this approximation. Two are resampling (bootstrap) methods, while the third is a Bayesian approach using Gibbs sampling.

5.6.1 Normal approximation assuming uncertainty from observed deaths only

The simplest approach is to only include the uncertainty from the observed deaths, that is, to assume that each predicted probability of death for observations in the unit of interest ($\hat{\pi}_i$) is estimated without error. Under this assumption a 100(1-α)% confidence interval for $\sum \hat{\pi}_i$ can be easily constructed. If, initially, the case is considered when the probability of death is constant for all infants in any particular unit (i.e. no case-mix adjustment), and that this probability follows a binomial distribution $\left(\sum_{i=1}^{n} d_i \sim B(\pi, n) \right)$, then the lower ($\pi_L$) and upper ($\pi_U$) limits for a 100(1-$\alpha$)% confidence interval are given by solution to the formulae (Armitage and Berry, 1994:120-121):

$$\frac{\sum_{i=1}^{n} d_i - n\pi_L - \frac{1}{2}}{\sqrt{\left(n\pi_L\left[1-\pi_L\right]\right)}} = z_{\alpha/2} \qquad (5.26)$$

and

$$\frac{\sum_{i=1}^{n} d_i - n\pi_U + \frac{1}{2}}{\sqrt{\left(n\pi_U\left[1-\pi_U\right]\right)}} = z_{100-\alpha/2} \qquad (5.27)$$

where: π_L is the lower limit of the confidence interval

π_U is the upper limit of the confidence interval

and the continuity correction ½ is included.

These can be rewritten as:

$$\frac{n\bar{d} - n\pi_L - \frac{1}{2}}{\sqrt{\left(n\pi_L\left[1-\pi_L\right]\right)}} = \frac{\bar{d} - \pi_L - \frac{1}{2n}}{\sqrt{\left(\frac{\pi_L\left[1-\pi_L\right]}{n}\right)}} = z_{\alpha/2} \qquad (5.28)$$

and

$$\frac{n\bar{d} - n\pi_U + \frac{1}{2}}{\sqrt{\left(n\pi_U\left[1-\pi_U\right]\right)}} = \frac{\bar{d} - \pi_L + \frac{1}{2n}}{\sqrt{\left(\frac{\pi_L\left[1-\pi_L\right]}{n}\right)}} = z_{100-\alpha/2} \qquad (5.29)$$

where: $\bar{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$

π_L and π_U can be obtained by rearranging the two equations above (still including the continuity correction) and finding the two solutions to π^* in the quadratic equation:

$$\frac{\left|\bar{d} - \pi^*\right| - \frac{1}{2n}}{\sqrt{\frac{\pi^*\left(1-\pi^*\right)}{n}}} = z_{\alpha/2} \qquad (5.30)$$

The two solutions are (details of the solution for the lower limit are given in Appendix C.3):

$$\pi_L = \frac{\left(2n\bar{d} + z^2_{\alpha/2} - 1\right) - z_{\alpha/2}\sqrt{z^2_{\alpha/2} - \left(2 + \left[1/n\right]\right) + 4\bar{d}\left(n\left[1-\bar{d}\right] + 1\right)}}{2\left(n + z^2_{\alpha/2}\right)} \qquad (5.31)$$

$$\pi_U = \frac{\left(2n\bar{d} + z^2_{\alpha/2} + 1\right) + z_{\alpha/2}\sqrt{z^2_{\alpha/2} + \left(2 - \left[1/n\right]\right) + 4\bar{d}\left(n\left[1-\bar{d}\right] - 1\right)}}{2\left(n + z^2_{\alpha/2}\right)} \qquad (5.32)$$

It is suggested that the continuity correction in the numerator (½) can be omitted if $n\pi_L$ or $n\left(1-\pi_U\right)$ are greater than 5 (Armitage and Berry, 1994:122). Without the continuity correction the solutions to (*5.30*) are:

$$\pi_L = \frac{\left(2n\bar{d} + z^2_{\alpha/2}\right) - z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4n\bar{d}\left(1-\bar{d}\right)}}{2\left(n + z^2_{\alpha/2}\right)} \qquad (5.33)$$

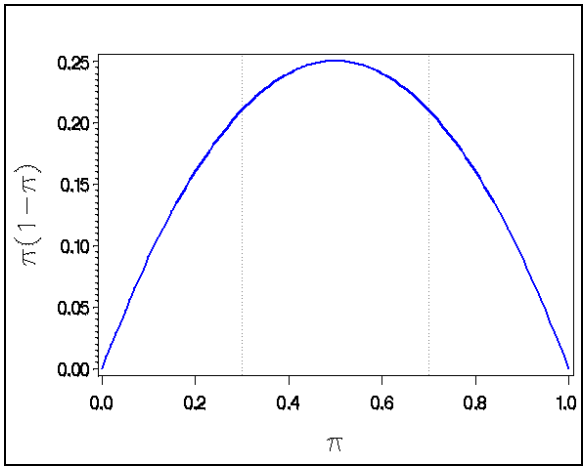
and

$$\pi_U = \frac{\left(2n\bar{d} + z^2_{\alpha/2}\right) + z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4n\bar{d}\left(1-\bar{d}\right)}}{2\left(n + z^2_{\alpha/2}\right)} \qquad (5.34)$$

Since these formulae are not particularly straightforward to use, $\pi_L\left[1-\pi_L\right]$ and $\pi_U\left[1-\pi_U\right]$ in (*5.26*) & (*5.27*) are usually replaced with $\hat{\pi}\left[1-\hat{\pi}\right]$, where $\hat{\pi}$ is the mortality rate from the

reference population. It is argued that this is an acceptable approximation because $\pi(1-\pi)$ changes slowly with changes in π. However, this is only true away from the limits of π, i.e. 0 and 1, as can be seen in Figure 5.10. Recommendations have been made to use this substitution only when $0.3 \le \pi \le 0.7$ and values of both $n\hat{\pi}$ and $n(1-\hat{\pi})$ equal at least 5 (Fleiss *et al,* 2003:29), or when both $n\hat{\pi}$ and $n(1-\hat{\pi})$ are equal to 10 or greater (Armitage and Berry, 1994:121). However, often such cautions are missing when they are reported in practice.

Figure 5.10 Plot of π(1-π) against π



With this substitution, (*5.33*) & (*5.34*) simplify to become the often-used approximations:

$$\pi_L = \bar{d} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} - \frac{1}{2n} \qquad (5.35)$$

and $$\pi_U = \bar{d} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} + \frac{1}{2n} \qquad (5.36)$$

although these are most often seen without the continuity correction $\frac{1}{2n}$: for example Rapoport *et al* (1994).

The approximations derived above apply when each observation has the same probability of the event (in this case death) occurring. Once adjustments are made for case-mix differences amongst the observations this is no longer the case. The approach often taken is to replace $\hat{\pi}$ with $\frac{\sum \hat{\pi}_i}{n}$ in (*5.35*) & (*5.36*) (Hosmer and Lemeshow, 1995), thus:

$$\pi_L = \bar{d} - z_{\alpha/2}\sqrt{\frac{\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)}{n^2}} - \frac{1}{2n} \qquad (5.37)$$

and $$\pi_U = \bar{d} + z_{\alpha/2}\sqrt{\frac{\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)}{n^2}} + \frac{1}{2n} \qquad (5.38)$$

With continuity correction

Following on, the limits of an estimated 100(1-α)% confidence interval for the standardized mortality ratio can then be seen to be given by:

$$SMR_L = \frac{\sum_{i=1}^{n} d_i - z_{\alpha/2}\sqrt{\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)} - \frac{1}{2}}{\sum_{i=1}^{n}\hat{\pi}_i} \qquad (5.39)$$

and $$SMR_U = \frac{\sum_{i=1}^{n} d_i + z_{\alpha/2}\sqrt{\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)} + \frac{1}{2}}{\sum_{i=1}^{n}\hat{\pi}_i} \qquad (5.40)$$

Without continuity correction

However, these are most often seen without the continuity correction:

$$SMR_L = \frac{\sum_{i=1}^{n} d_i - z_{\alpha/2}\sqrt{\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)}}{\sum_{i=1}^{n}\hat{\pi}_i} \qquad (5.41)$$

and $$SMR_U = \frac{\sum_{i=1}^{n} d_i + z_{\alpha/2}\sqrt{\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)}}{\sum_{i=1}^{n}\hat{\pi}_i} \qquad (5.42)$$

'Full' method

However, (*5.41*) & (*5.42*) are derived using the assumption that it is appropriate to replaced $\pi_L[1-\pi_L]$ and $\pi_U[1-\pi_U]$ with $\hat{\pi}[1-\hat{\pi}]$. If this is not the case then, in the situation where $\hat{\pi}_i$ is not fixed, the two limits are given by:

$$SMR_L = \frac{n.\pi_L}{\sum_{i=1}^{n} \hat{\pi}_i} \qquad (5.43)$$

and

$$SMR_U = \frac{n.\pi_U}{\sum_{i=1}^{n} \hat{\pi}_i} \qquad (5.44)$$

where π_L and π_U are given by (*5.31*) & (*5.32*).

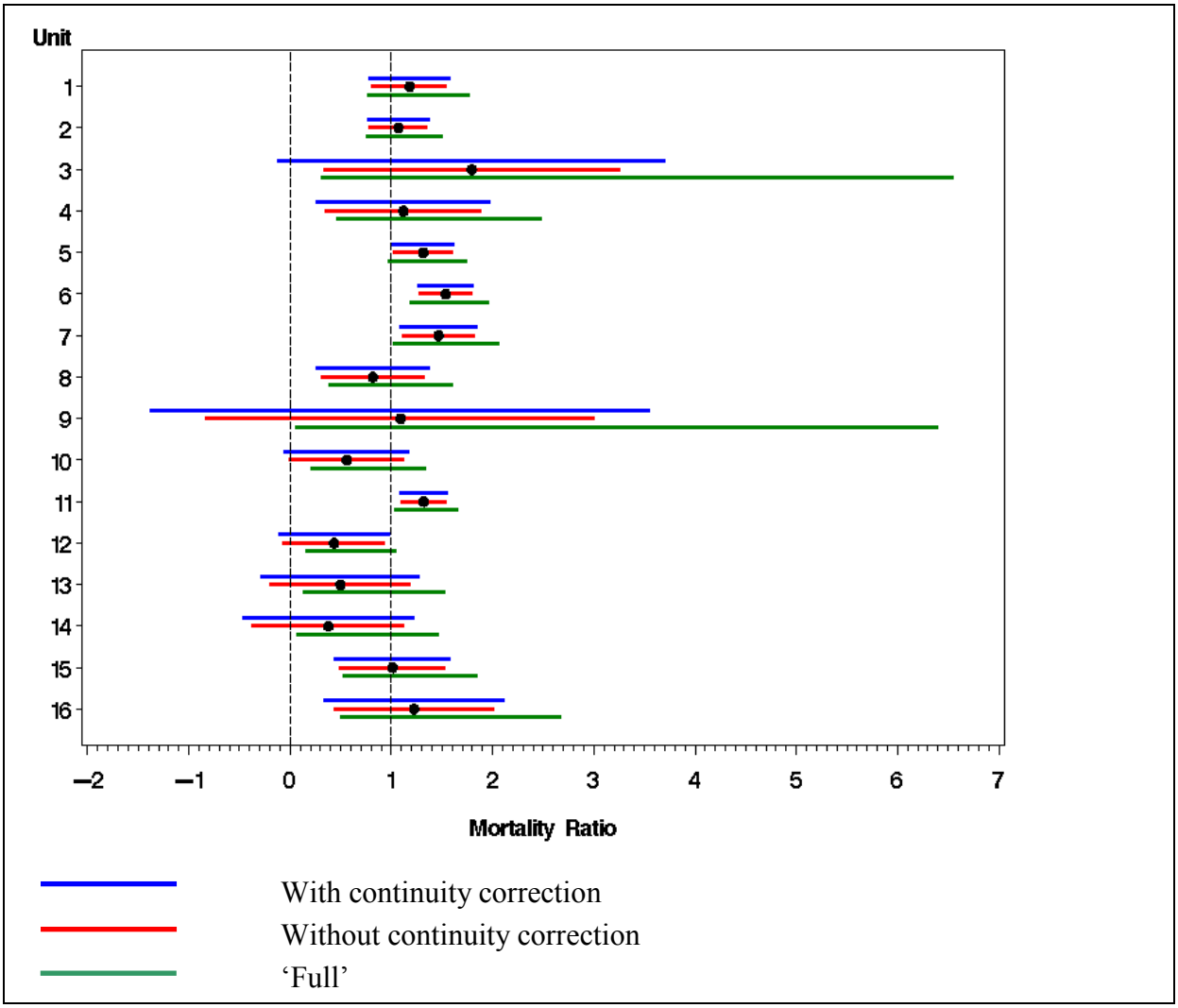
Comparison of methods based on the Normal approximation

In this Section three approaches to the estimation of the limits for confidence intervals of the SMR have been proposed: (*5.39*) & (*5.40*) (referred to as 'With CC'), (*5.41*) & (*5.42*) ('Without CC') and (*5.43*) & (*5.44*) ('Full'). The estimated limits when these methods are applied to the TNS data are shown in Table 5.16 and Figure 5.11.

Table 5.16 95% confidence intervals for SMR using Normal approximation methods for TNS data

Unit	SMR	95% Confidence Interval		
		With CC	Without CC	'Full'
1	1.18	0.78 to 1.59	0.81 to 1.56	0.76 to 1.79
2	1.07	0.76 to 1.38	0.78 to 1.37	0.75 to 1.51
3	1.80	-0.12 to 3.71	0.33 to 3.26	0.31 to 6.56
4	1.12	0.26 to 1.99	0.35 to 1.90	0.46 to 2.49
5	1.32	1.01 to 1.63	1.02 to 1.61	0.97 to 1.76
6	1.54	1.26 to 1.82	1.27 to 1.81	1.19 to 1.97
7	1.47	1.08 to 1.86	1.11 to 1.83	1.02 to 2.08
8	0.82	0.26 to 1.38	0.31 to 1.33	0.39 to 1.62
9	1.09	-1.38 to 3.57	-0.83 to 3.02	0.06 to 6.40
10	0.56	-0.07 to 1.19	-0.01 to 1.13	0.21 to 1.35
11	1.32	1.08 to 1.56	1.09 to 1.55	1.04 to 1.67
12	0.44	-0.12 to 0.99	-0.07 to 0.95	0.16 to 1.06
13	0.50	-0.29 to 1.29	-0.20 to 1.20	0.13 to 1.55
14	0.38	-0.47 to 1.23	-0.37 to 1.14	0.07 to 1.47
15	1.01	0.44 to 1.59	0.49 to 1.54	0.52 to 1.85
16	1.23	0.33 to 2.12	0.43 to 2.02	0.50 to 2.69

Figure 5.11 95% confidence intervals for SMR using Normal approximation methods for TNS data



The two methods using the substitutions $\pi_L[1-\pi_L]=\hat{\pi}[1-\hat{\pi}]$ and $\pi_U[1-\pi_U]=\hat{\pi}[1-\hat{\pi}]$ (blue and green) both produce intervals that are symmetrical about the point estimate for the SMR. This can lead to lower limits less that zero. In addition, the intervals calculated without the continuity correction are narrower than those calculated with it. The intervals calculated using the 'full' method are not symmetrical and tend to have higher lower and upper limits than those obtained using the other two methods. The differences between the methods are particularly noticeable for the small units. Although more complex to calculate, the intervals calculated using the 'Full' method have a more rigorous theoretical base, are not restricted to lie symmetrically about the point estimate and do not include (implausible) negative values. These would, therefore, seem to be the more appropriate choice. However, the coverage properties of confidence intervals calculated using these methods will be further investigated later in this Chapter (§5.7).

None of these methods take into account any uncertainty in the expected number of deaths. Two possible extensions to the Normal approximation are illustrated next. The first was proposed by Hosmer & Lemeshow (1995) and the second by Zhou & Romano (1997).

5.6.2 Extensions to the Normal approximation method

Although the methods outlined above are widely used, one problem is that it is assumed that uncertainty only arises from the number of observed deaths. However, uncertainty can arise from two sources. First, as has been included above, the observed number of deaths is an observation from a random process and, therefore, has uncertainty associated with it. The second source of variability arises from the uncertainty in the estimates of the model parameters used to estimate the expected number of deaths. Although the variability around the observed deaths is likely to be far greater that that from the expected, as the model providing these estimates will have a larger number of observations, ignoring either source of uncertainty will produce intervals that are too narrow (Signorini and Weir, 1999). The uncertainty referred to in this context only refers to the sampling uncertainty and does not include any uncertainty from model mispecification.

Hosmer & Lemeshow (1995)

Taking the logarithm of the SMR gives:

$$\log(O/E) = \log(O) - \log(E)$$

and using the delta method to approximate the variance of $\log(O/E)$, assuming independence between O and E gives:

$$\hat{Var}\left[\log\left(\%_E\right)\right] = \frac{\hat{Var}(O)}{o^2} + \frac{\hat{Var}(E)}{\hat{e}^2} \qquad (5.45)$$

The estimated variance of the expected total is given by:

$$Var\left(\hat{E}\right) = \mathbf{1}'\hat{\mathbf{V}}_j\mathbf{X}_j\left(\mathbf{X}'_{\mathbf{R}(-j)}\hat{\mathbf{V}}_{\mathbf{R}(-j)}\mathbf{X}_{\mathbf{R}(-j)}\right)^{-1}\mathbf{X}'_j\hat{\mathbf{V}}_j\mathbf{1} \qquad (5.46)$$

where: subscript j refers to the unit of interest;

subscript $\mathbf{R(-}j\mathbf{)}$ refers to the units other than the unit of interest;

$\mathbf{V}$ is the m x m diagonal matrix: $diag\left\{\hat{\pi}_i\left(1-\hat{\pi}_i\right)\right\}$;

$\mathbf{X}$ is the data matrix

$\mathbf{1}$ is the m x m diagonal matrix: $diag\{1\}$.

In practice this is simplified by recognising that:

$$\left(\mathbf{X}'_{\mathbf{R}(-j)}\hat{\mathbf{V}}_{\mathbf{R}(-j)}\mathbf{X}_{\mathbf{R}(-j)}\right) = \mathbf{S}\left(\hat{\boldsymbol{\beta}}_{\mathbf{R}(-j)}\right) \qquad (5.47)$$

that is, the asymptotic covariance matrix of the parameter estimates $\hat{\boldsymbol{\beta}}_{\mathbf{R}(-j)}$ (i.e. the inverse of the information matrix). This enables the variance of the expected value to be estimated either where the raw data are available, as in this case, or where the covariance matrix is available from the logistic regression model. The latter possibility may be useful where the function to estimate the expected values has been developed by one group and the original data are not available to others.

Hosmer and Lemeshow took the estimated variance of the observed number of deaths to be:

$$\hat{Var}(O) = \frac{\sum \hat{\pi}_i(1-\hat{\pi}_i)}{n} \qquad (5.48)$$

or, in matrix notation:

$$\hat{Var}(O) = \mathbf{1}'\hat{\mathbf{V}}_j\mathbf{1} \qquad (5.49)$$

These can then be used with (*5.45*) to estimate a confidence interval for the SMR:

$$\left[\left(e^{\log(\%_e)-z_{1-\alpha/2}\sqrt{\hat{Var}(\log(\%_E))}}\right) \text{ to } \left(e^{\log(\%_e)-z_{1-\alpha/2}\sqrt{\hat{Var}(\log(\%_E))}}\right)\right] \qquad (5.50)$$

The authors argued that this estimate is appropriate as the logarithm of the ratio is more likely to be approximately Normally distributed than the ratio itself (Hosmer and Lemeshow, 1997). Using the logarithm of the ratio also has the advantage that the lower end point will always be positive. One characteristic noted with this estimate is that, with the examples shown in the original paper (Hosmer and Lemeshow, 1995), the lower limit of this interval was always greater than the lower limit of the interval using var(*O*) alone. This was particularly noticeable for small datasets and is counterintuitive as including more uncertainty should reduce the value of the lower limit (and raise the value of the upper). This was claimed to be a 'fault' with the method (Zhou and Romano, 1997). However, this effect is an artefact of moving from the (inappropriate) linear scale to the (more appropriate) logarithmic scale.

Zhou & Romano (1997)

After pointing out the characteristic of Hosmer & Lemeshow's method described above, Zhou & Romano (1997) proposed an extension to the method to return the interval to the linear

scale. The derivation of their method is the same as that given by Hosmer and Lemeshow, up to the point of using the delta method to obtain an estimate of *var*(*E*). The delta method is then applied again to obtain an estimate for the variance of the exponential of the logarithm of the ratio, i.e. the ratio itself:

$$\hat{\text{var}}\left(e^{\log[O/E]}\right) = \left(e^{\log[O/e]}\right)^2 \left(\frac{\hat{\text{var}}[O]}{o^2} + \frac{\hat{\text{var}}[E]}{e^2}\right) \qquad (5.51)$$

Hence:
$$\hat{\text{var}}\left(\frac{O}{E}\right) = \left(\frac{o}{e}\right)^2 \left(\frac{\hat{\text{var}}[O]}{o^2} + \frac{\hat{\text{var}}[E]}{e^2}\right) \qquad (5.52)$$

This estimate can then be used to create a confidence interval for the ratio:

$$\left(\left[O/E\right] - z_{1-\alpha/2}\left[O/E\right]\sqrt{\frac{\hat{\text{var}}[O]}{o^2} + \frac{\hat{\text{var}}[E]}{\hat{e}^2}}\right) \textbf{\textit{ to }} \left(\left[O/E\right] + z_{1-\alpha/2}\left[O/E\right]\sqrt{\frac{\hat{\text{var}}[O]}{o^2} + \frac{\hat{\text{var}}[E]}{\hat{e}^2}}\right) \qquad (5.53)$$

The most obvious problem with this interval is that the lower limit can be negative. However, the authors believe that their method is superior as it overcomes the 'problem' with the lower confidence limit found with the Hosmer & Lemeshow method.

5.6.3 Bootstrap

All of the methods described so far are based on the Normal approximation. However, the assumptions made may not hold, especially for small samples. An alternative approach is to use bootstrap methods. In this thesis two approaches using bootstrap methods to estimate confidence intervals for the SMR were investigated.

Accounting for uncertainty of observed deaths only

The first approach involved fitting the appropriate logistic regression model and then repeated sampling with replacement from the observations of the unit of interest. The size of each sample was equal to the number of observations in that unit. For each sample the SMR was calculated using the observed and expected number of deaths in that sample. The distribution of the estimated SMR can then be used to obtain confidence limits. This can be done in several ways. A naïve approach (that used in §3.3.1) is to use the percentile method and to simply report the observed 2.5th and 97.5th percentiles as the limits of the 95% confidence interval. However, this method is known not to work well in small samples, especially when the distribution is asymmetrical (Chernick, 1999:53; Davison and Hinkley, 1997:203; Good, 1999:91), and this is likely to be the case for the SMR. Alternative bootstrap methods have

been proposed that may be more appropriate, but this thesis does not aim to discuss them all. Rather, one method that has previously been suggested for the SMR will be illustrated (Hosmer and Lemeshow, 1995).

Confidence limits can be found for a statistic such as the SMR using the **bias-corrected and accelerated bootstrap** method (BCa). It is assumed that there is a monotone transformation h such that $\phi = h(\theta)$, where θ is the parameter of interest (here the SMR) and ϕ approximately follows a Normally distribution with mean $\phi - z_0\tau_\phi$. In this expression z_0 is the bias correction and takes the value of the difference between the point estimate of the parameter of interest (i.e. $\hat{\theta}$) and the value of the 50th percentile from the bootstrap samples ($\hat{\theta}^*_{50}$): therefore $z_0 = \hat{\theta} - \hat{\theta}^*_{50}$. The term τ_ϕ represents the standard deviation of $\hat{\phi}$ and depends on θ given that $\tau_\phi = 1 + a\phi$, where a is the acceleration constant often defined as one-sixth of the skewness and is usually estimated from the data (Efron and Tibshirani, 1993:186; Chernick, 1999:53; Davison and Hinkley, 1997:203).

Accounting for uncertainty of observed and expected deaths

While useful, the method above does not take into account the full uncertainty of the model since only the observations from the unit of interest are sampled. An alternative approach is to sample from the whole population, with the sample size for each unit equal to its number of observations, and then to fit a model to each sample. An estimated SMR can be calculated for each bootstrap sample and confidence intervals calculated as described above. This approach would ensure that the uncertainty in the model parameter estimates was included. However, since a logistic regression model is fitted for each bootstrap sample, this is more computationally intensive than the first bootstrap method outlined above. In this thesis, such an approach was either impracticable (§5.7.1) or other problems were found with the whole approach (§5.8.5). Therefore, simple percentile confidence limits were used with this second bootstrap method.

5.6.4 Bayesian Method

An alternative approach is to use Bayesian methods. One suitable model is, for Unit *j*:

$$\hat{SMR}_j = \frac{\sum_{i=1}^{n_j} \hat{d}_i}{\sum_{k=1}^{n_j} \hat{\pi}_i} \qquad (5.54)$$

where: $\hat{d}_i = \frac{1}{1+e^{\left(-\hat{\boldsymbol{\beta}}_j \mathbf{X}_i\right)}}$

$$\hat{\pi}_i = \frac{1}{1+e^{\left(-\hat{\boldsymbol{\beta}}_R \mathbf{X}_i\right)}}$$

where: $\hat{\boldsymbol{\beta}}_j$ are the parameter estimates obtained from the observations from Unit *j*

$\hat{\boldsymbol{\beta}}_R$ are the parameter estimates obtained from the observations the reference population

This approach involves parameter estimates from two logistic regression models. The denominator is derived as before, from logistic regression parameter estimates obtained using the reference data. From these, indirectly standardized probabilities of death are estimated for each observation from the unit of interest and these are then summed to obtain the total expected number of deaths. However, with this method the value of the numerator is obtained using a separate logistic regression model (but with the same risk-adjustment variables) in which the parameters are estimated using the observations from the unit of interest. The parameter estimates from this second model are then used to estimate a second predicted probability for each observation (i.e. $\hat{d}_i$). These probabilities are then summed to obtain an estimate of the 'observed' number of deaths. The ratio of these two summations is then taken as an estimate of the SMR. This process is repeated at each iteration to obtain a posterior distribution for the SMR from which credible intervals can be estimated (§3.2.2). An obvious difficulty with this approach is that the parameter estimates derived using the observations from the unit of interest (i.e. $\hat{\boldsymbol{\beta}}_j$) are likely to be poorly estimated. For small units the prior distributions chosen are likely to dominate. Some previous work has ignored this source of uncertainty and has only included the uncertainty in the denominator (Austin *et al,* 2001; Austin, 2002). However, such an approach is likely to seriously underestimate the true uncertainty and produce confidence intervals that are too narrow.

5.7 Simulation study

It is unclear which method for estimating confidence intervals for SMRs is the most appropriate. Of particular interested in this thesis is their application to small units.

5.7.1 Methods of simulation study

The performance of the estimation methods described previously in this Section (§5.6.1 to §5.6.4) was investigated using a simulation study. For each simulation, two datasets were created: a small dataset to represent the health care provider of interest and a large set to represent the reference data. Using the same empirical distribution for each dataset, morbidity scores were input and observations were sampled using a logistic model with a known linear predictor. These observations then formed the basis for estimation of the outcome ratio and confidence intervals. Twenty-seven different scenarios were simulated with 1,000 repetitions in each case, varying: (i) the size of the dataset of interest; (ii) the size of the reference dataset; (iii) the underlying probability of death within the dataset of interest. For each scenario the Type I error rates and coverage were calculated and compared. The choice of 1,000 repetitions was a pragmatic decision that hoped to balance precision with practicality.

The size of the target dataset was set at 50, 100 and 200 observations and the reference data set having 500, 1000 and 2000 observations to try to represent realistic sample sizes in neonatal intensive care.

For each observation the probability of an event was calculated using the logistic transformation and a known linear predictor, i.e.:

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \beta_S . score)}}$$

where β_0 and β_S are specified and *score* represents a morbidity severity score which had the same given empirical distribution in each data set:

Score	1	2	3	4	5	6	7	8	9	10
Proportion	0.14	0.18	0.20	0.16	0.10	0.08	0.06	0.04	0.02	0.02

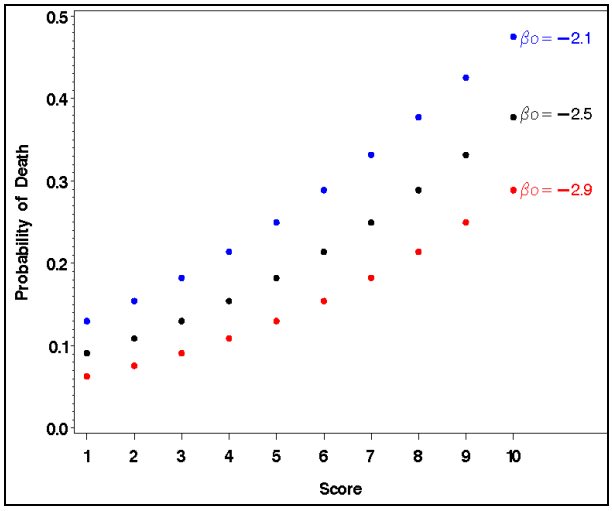
In all cases $\beta_S = 0.2$ but β_0 was varied. Three values for β_0 were chosen:

- $\beta_0 = -2.5$ was used for the reference data in all scenarios. It was also used for the target data to simulate no underlying differences between the populations;

- β_0 = -2.1 was used for the target data to approximate an Odds Ratio for mortality between the two populations of ¾ (mortality ratio of approximately 1.37);
- β_0 = -2.9 was used for the target data to approximate an Odds Ratio for mortality between the two populations of ⅔ (mortality ratio of approximately 0.71).

The probability of death for each given value of β_0 at each value of 'score' is shown in Figure 5.12.

Figure 5.12 Probability of death by score



The distribution of 'Score' was selected to mimic the observed skewed distribution of scores such as CRIB in neonatal populations. In addition, a relatively high probability of death was simulated to try to avoid samples where there were no observed deaths. The methods proposed by Hosmer & Lemeshow and Zhou & Romano require division by the number of observed deaths, therefore the limits are undefined when the observed number is zero. The methods based on the bootstrap are also unsuitable for units with no observed events. An alternative approach, for the methods based on the Normal approximation, would have been to add a small constant to all zero observations, but it was felt that this would make interpretation more difficult as the effect of adding a constant is unknown. A further solution would have been to discard all simulations with zero observed deaths, but this would produce a biased set of intervals. This is discussed further in §0.

For each observation an event (Y_i) was simulated where:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

The simulated values Y_i, for the reference part of the simulated data set, were then used in a new logistic regression model to estimate β_0^* and β_S^*, where $logit(Y_i) = \beta_0^* + \beta_S^* .score$. In the Bayesian approach a similar model was estimated for the simulated data of the unit of interest. Confidence (credible) intervals were then estimated using the methods previously described and the proportion of intervals not containing the true ratio reported. In total eight methods were investigated, including the three methods derived from the Normal approximation discussed in §5.6.1.

For both bootstrap methods 1,000 replications were used, a figure thought to be generally *"safe"* for estimation (Davison and Hinkley, 1997:156). While the bias-corrected and accelerated bootstrap method (BC_a) was used for the simpler approach of bootstrapping the predicted probabilities from a single model, such an approach was not possible for fitting a logistic regression model to each replication. Computer memory limitations meant that this approach was not possible without the use of loops in the programme, leading to 1,000 bootstrap loops within 1,000 datasets (i.e. estimating 1 million logistic regression models) and then using jackknife methods to estimate the value of the acceleration constant *a*. All of this meant that each scenario would take an extremely long time to run. Because of this problem, percentile bootstrap intervals were estimated but, even so, due to the loops, each scenario took up to two days to run.

For the Bayesian approach a 1,000 iteration 'burn-in' was used and then 10,000 iterations for estimation. The prior distribution chosen for all of the parameter estimates was $Normal(0,1000^2)$

The SAS macros and WinBUGS code used to simulate the data and to calculate the intervals are given in Appendix F.1.

5.7.2 Results

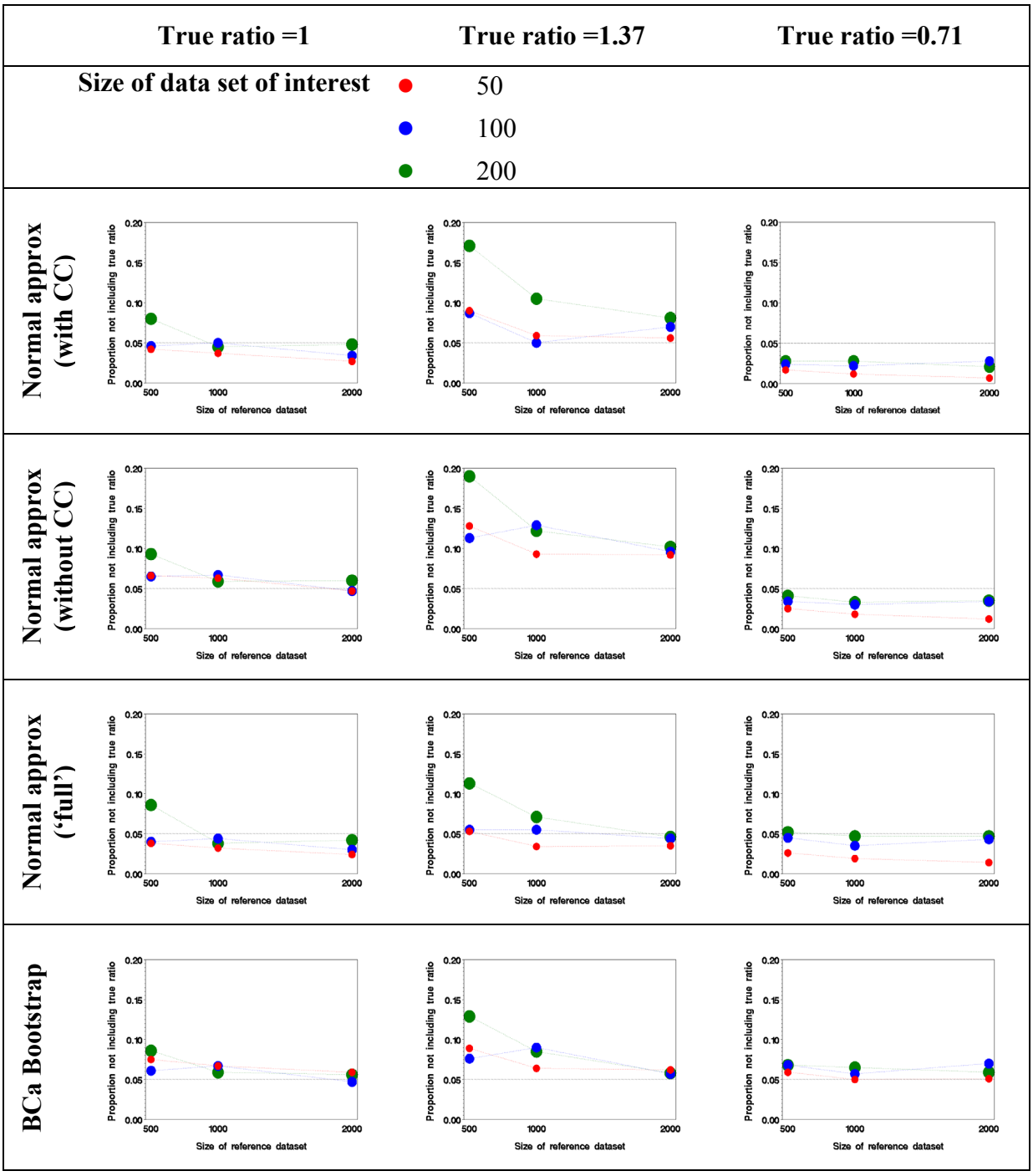
The mean event rate and range for the 1,000 data sets in each scenario are given in Appendix F.2. The mean event rates were consistent across all of the data sets, although for two scenarios there were some data sets with no events. For these two scenarios only the non-zero event data sets were included in any further analyses. This was unlikely to affect the results described below.

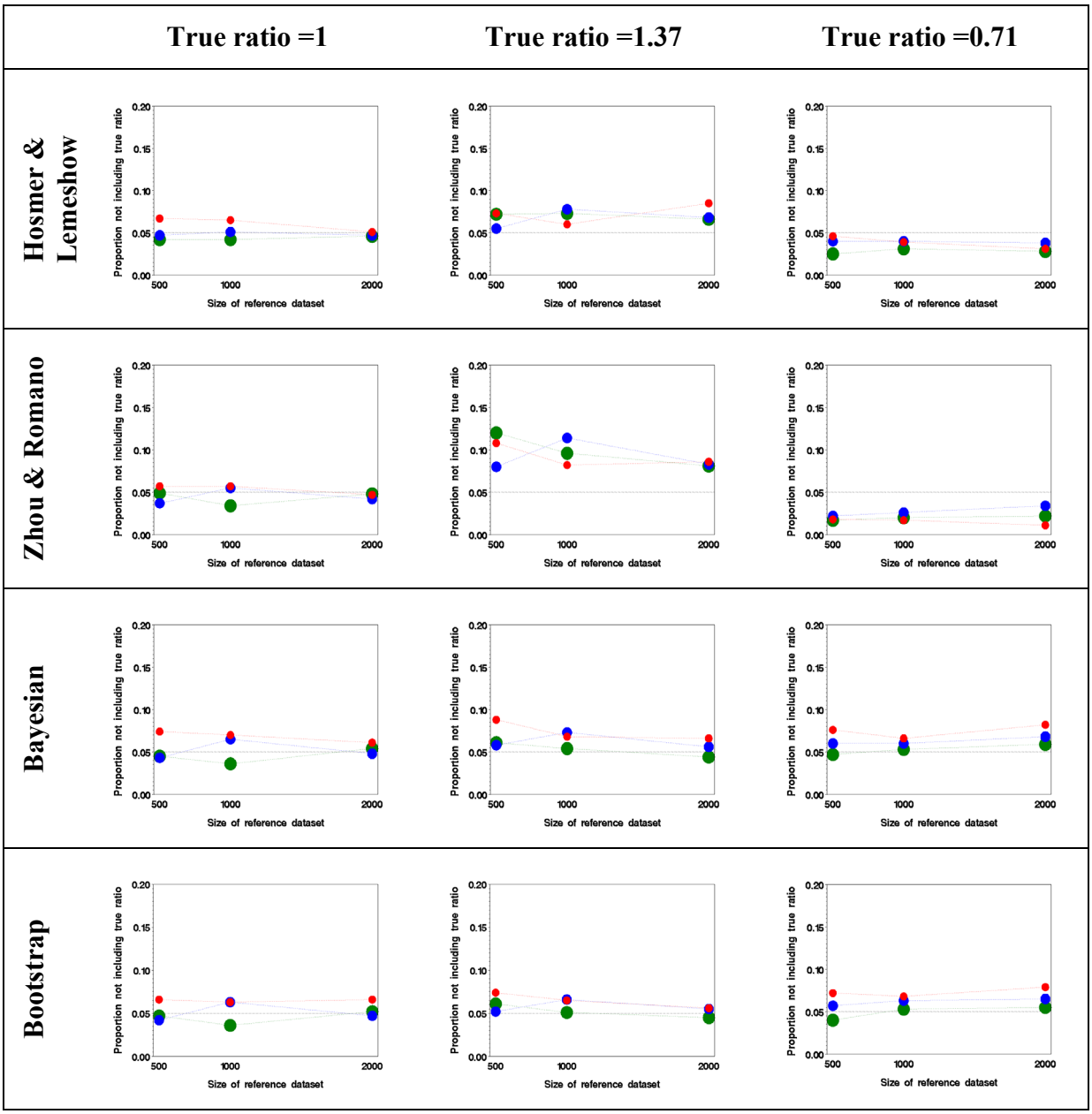
Histograms and Normal and log-Normal distributions plots for the simulated SMRs are also shown in Appendix F.2. Inspection of these plots offers some evidence that the distribution

of these ratios more closely approximates a log-Normal distribution than the Normal distribution. This is not unexpected, as previously discussed.

Each estimated 95% confidence interval was inspected to see whether it contained the value of the true SMR and the observed coverage of the estimated intervals of the 27 scenarios are shown in Figure 5.13. Further details are given in Appendix F.3.

Figure 5.13 Observed coverage of estimated 95% confidence intervals for SMR





In general, the coverage rates are closer to the true rate of 95% when the true SMR for the simulated data is equal to one. For most of the methods the coverage rate was too small when the SMR was greater than one, and too large when less than one. The bootstrap and Bayesian methods were the exceptions to this with coverage rates generally (slightly) less than 95% when the true ratio was 0.71.

The upper limits of the estimated intervals, when the true ratio was unity and the sample sizes were 100 and 1,000, are shown in Figure 5.14 and the lower limits in Figure 5.15. For simplicity, only the most commonly used Normal approximation method is illustrated ('without CC'). Similar plots for the other scenarios are in Appendix F.3.

Figure 5.14 Upper limits of 95% confidence (credible) intervals

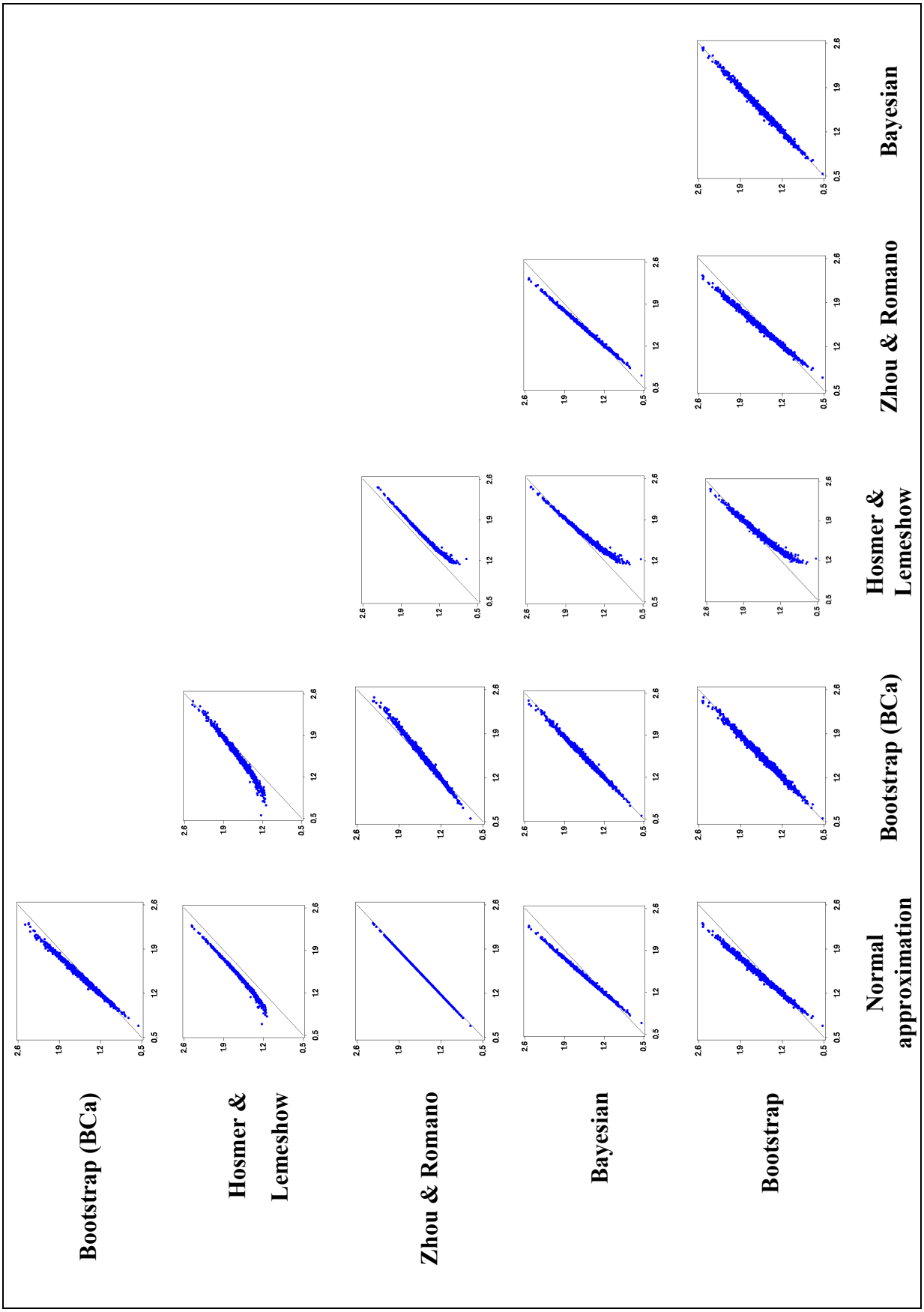
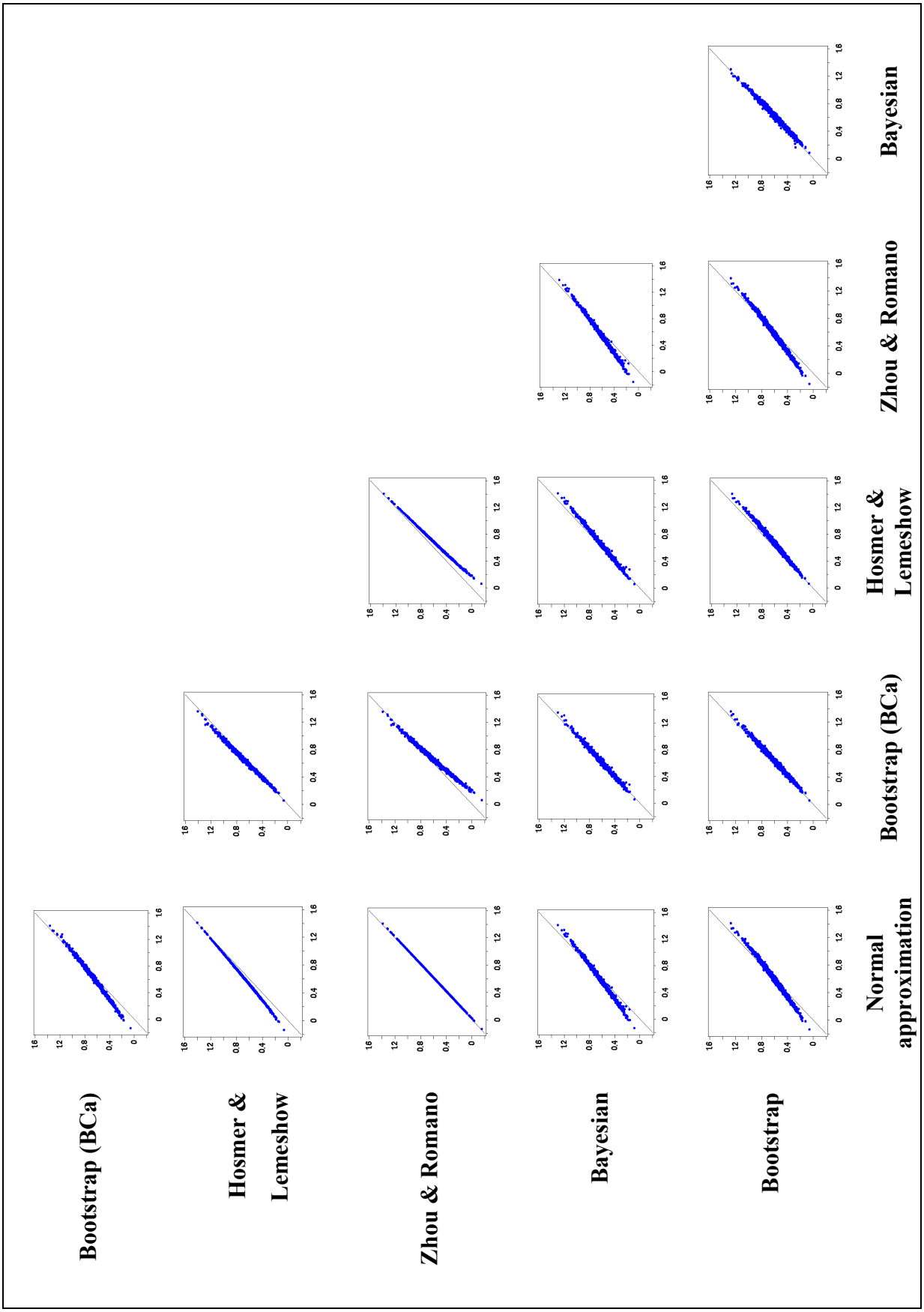


Figure 5.15 Lower limits of 95% confidence (credible) intervals



These plots show the negative lower limits from the normal approximation method and the method proposed by Zhou & Romano. There was little difference between the values estimated by these two methods, with the intervals estimated by the latter method being slightly wider, reflecting the addition of the uncertainty in the expected number of events. In addition, there was little difference between estimates from the bootstrap and Bayesian approaches. The method proposed by Hosmer & Lemeshow tended to estimate very high values for the upper limits and rarely produce intervals where the value of the upper limit was less than the true value for the SMR. The influence of the choice of prior distributions in the Bayesian analysis is unknown but, for the small units at least, is likely to be influential (this is investigated further in §5.8). Therefore, the final bootstrap method, although computationally intensive, appeared to offer the most appropriate approach at this stage.

Although informative, the simulation study described here may be an oversimplification of the situation found in a 'real' study. The models in the simulation study are correctly specified, with all covariates included in the correct form. This is unlikely to be the case with the TNS data analysed in this thesis. It is of interest, therefore, to compare the estimated confidence intervals when these methods are applied to the TNS data. This is described next.

5.8 Application of confidence interval estimation methods to TNS data

The previous Section showed differences in the coverage properties of the confidence intervals estimated using different methods. Six of the methods outlined above are illustrated using the data from TNS. As previously, gestational age at birth was included in the model and the 'deviation from the mean' parameterisation used: i.e. a separate model was fitted for each unit to estimate the mean log odds of death in the other 15 units (see §5.4 for more details):

$$\text{logit}(\pi_i) = g_i = \beta_0 + \beta_G . gest_i + \sum_{k=2}^{16} \beta_k I_k \qquad (5.55)$$

5.8.1 Normal approximation

Methods based on the Normal approximation have been illustrated using TNS data in §5.6.1. In this Section the results using the most commonly used approach, Normal approximation

without the continuity correction (*5.41*) & (*5.42*), was repeated for comparison with the alternative methods.

5.8.2 Hosmer & Lemeshow and Zhou & Romano

These two extensions to the Normal approximation method were both applied generally as described in §5.6.2. However, with the TNS data the model used for the reference data contained indicator variables representing the NICUs (*5.14*). Obviously, these indicators variables did not exist in the design matrix for the unit of interest ($\mathbf{X}_j$ in 5.46). Therefore, only the components in the covariance matrix $\hat{\boldsymbol{\beta}}_{\mathbf{R}(-j)}$ that related to the intercept and gestational age were used to estimate the variance (*5.46*). The likely effect of this was to reduce the estimate of $\hat{Var}(E)$ but the size of this reduction is not known, and was not investigated in this thesis.

5.8.3 Bootstrap

The two bootstrap methods were applied in an identical way to the simulation study (§5.6.3): a BCa bootstrap interval was estimated for the first method and a percentile interval of the second. One thousand bootstrap samples were simulated for each method.

5.8.4 Bayesian analysis

This thesis aims to illustrate the different methods of estimating the confidence intervals rather than to be an examination of the effects of various prior distributions. However, it is recognised that the choice of prior distribution can influence the estimates of the confidence intervals, particularly for the small units.

Using the Bayesian approach shown in *(5.56)*, prior probability distributions were required to be specified for the intercepts β_0 and β_{R0}, the regression parameters for gestational age β_G and β_{RG} and for the indicator parameters for the reference units $\boldsymbol{\beta}_K$:

$$\hat{SMR}_j = \frac{\sum_{i=1}^{n_j} \hat{d}_i}{\sum_{k=1}^{n_j} \hat{\pi}_i} \qquad (5.56)$$

where:
$$\hat{d}_i = \frac{1}{1+e^{-\left(\hat{\beta}_0+\hat{\beta}_G.gest_i\right)}}$$
$$\hat{\pi}_i = \frac{1}{1+e^{-\left(\hat{\beta}_{R0}+\hat{\beta}_{RG}.gest_i\right)}}$$

The parameter estimates $\hat{\beta}_{RO}$ and $\hat{\beta}_{RO}$ are obtained using the data from the other 15 units:

$$\hat{g}_i = \hat{\beta}_{R0} + \hat{\beta}_{RG}.gest_i + \sum_{\substack{k=1 \\ k \neq j}}^{16} \hat{\beta}_k.I_k$$

Where the indicator variables I_k follow the deviation parameterization described in §5.3.1.

Previously, in §3.4.2, a prior distribution was specified for the underlying probability of death (π_0 and π_{R0}). However, it is more straightforward to specify a prior distribution for *logit*(π), i.e. β_0 and β_{R0}, although this is likely to be less intuitive to interprete. In this section both approaches were undertaken: three probability distributions were specified for π_0 and π_{R0}, and four for β_0 and β_{R0} (Figure 5.16). As the value of gestational age included in the model was centred on the median value (30 weeks), this was the value at which the intercept, the mean log odds of deaths for the other 15 units, was estimated. There was some evidence from the analysis of previous TNS data that the expected in-unit mortality rate for infants born at 30 weeks gestational age was about 5% (Draper *et al,* 1999). These seven probability distributions allowed an investigation into the influence of the location and precision of the prior distributions (Table 5.17) on the parameter estimates.

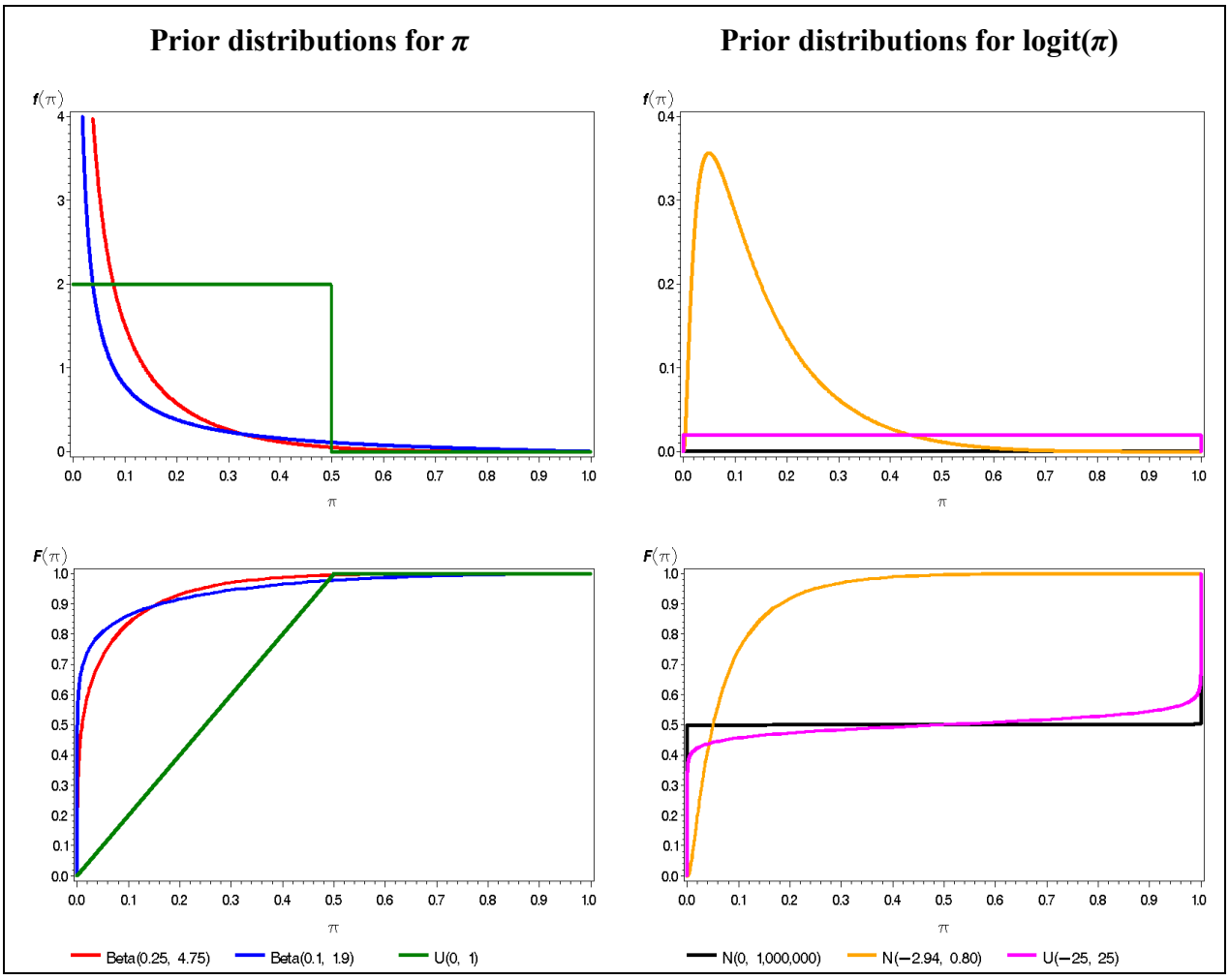
Table 5.17 Prior distributions

Distribution	***E*(π)**	***Var*(π)**	**P(π<*x*)≈0.025**	**P(π>*x*)≈0.025**
Prior on *π*				
Uniform(0,1)	0.50	0.083	$x = 0.025$	$x = 0.975$
Beta(0.25, 4.75)	0.05	0.008	$x = 6\text{x}10^{-7}$	$x = 0.32$
Beta(0.1, 1.9)	0.05	0.016	$x = 4\text{x}10^{-17}$	$x = 0.48$
Prior on *logit(π)* [i]				
Normal(0, 1000^2)	0.50	250.00	$x < 1\text{x}10^{-99}$	$x > 0.99999$
Normal(-2.94, 1000^2)	0.05	47.69	$x < 1\text{x}10^{-99}$	$x > 0.99999$
Normal(-2.94, 0.80)	0.05	0.043	$x = 0.009$	$x = 0.22$
Uniform(-25, 25)	0.50	11.51	$x = 5\text{x}10^{-11}$	$x > 0.99999$

[i] Using the delta method: $\text{Var}(\pi) = \left(\frac{e^{\text{logit}[\pi]}}{\{1 + e^{\text{logit}[\pi]}\}^2} \right)^2 .Var(\text{logit}[\pi])$

The probability density functions (PDFs) and cumulative density functions (CDFs) for these probability distributions are shown in Figure 5.16.

Figure 5.16 PDFs and CDFs for prior distrbutions



Ten scenarios were selected to investigate the influence of the choice of prior distributions on the parameter estimates. The first seven looked at the effect of various prior distributions for β_0 and β_{R0}. For the next scenario the estimated values for β_{R0} and β_{RG} were given as the mean values for Normal distributions with large variances to be used as prior distributions for β_0 and β_G. For the final two scenarios alternative probability distributions were specified for $\boldsymbol{\beta}_K$.

These approaches to the specification of the prior probabilities are first illustrated using Unit 16 (Table 5.18). Equivalent tables for Units 3 and 8 are given in Appendix F.4 for further information. A 1,000 iteration burn-in was inspected (§3.2.2) to ensure that sampling occurred from the whole of the target distribution, plots for Unit 8 are given in Appendix F.4 as an example. Once it had been confirmed that samples appeared to be being taken from the full target distribution, 10,000 further samples were then taken.

Table 5.18 Parameter estimates using various prior distributions (Unit 16)

Prior distibution					**Parameter Estimate**							
β_{R0}	β_{RG}	β_K	β_0	β_G	$\hat{\beta}_{R0}$	$\hat{\beta}_{RG}$	$\hat{\beta}_0$	$\hat{\beta}_G$	$\Sigma\hat{d}_i$	$\Sigma\hat{\pi}_i$	$\hat{\text{SMR}}$	(95% CI)
Frequentist model					-3.95	-0.65	-3.45	-0.55	6.00	4.88	1.23	
$\pi \sim U(0,1)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim U(0,1)$	$N(0, 1000^2)$	-4.03	-0.66	-3.24	-0.50	6.70	4.63	1.46	(0.64 to 2.74)
$\pi \sim B(0.25, 4.75)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim B(0.25, 4.75)$	$N(0, 1000^2)$	-4.08	-0.66	-3.54	-0.57	5.89	4.49	1.31	(0.54 to 2.55)
$\pi \sim B(0.1, 1.9)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim B(0.1, 1.9)$	$N(0, 1000^2)$	-4.08	-0.66	-3.57	-0.57	5.87	4.51	1.31	(0.54 to 2.58)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.07	-0.66	-3.63	-0.59	5.73	4.52	1.28	(0.52 to 2.49)
$N(-2.94, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(-2.94, 1000^2)$	$N(0, 1000^2)$	-4.07	-0.66	-3.63	-0.59	5.73	4.52	1.28	(0.52 to 2.49)
$N(-2.94, 0.80)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(-2.94, 0.80)$	$N(0, 1000^2)$	-4.05	-0.66	-3.41	-0.54	6.23	4.58	1.37	(0.60 to 2.57)
$U(-25, 25)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$U(-25, 25)$	$N(0, 1000^2)$	-4.08	-0.66	-3.64	-0.59	5.75	4.50	1.28	(0.53 to 2.54)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(\beta_{R0}, 1000^2)$	$N(\beta_{RG}, 1000^2)$	-4.07	-0.66	-3.63	-0.59	5.73	4.52	1.28	(0.52 to 2.49)
$N(0, 1000^2)$	$N(0, 1000^2)$	$U(-2, 2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.05	-0.66	-3.62	-0.59	5.75	4.60	1.25	(0.50 to 2.45)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.00	-0.66	-3.61	-0.58	5.80	4.78	1.22	(0.50 to 2.36)

Potential difficulties with the estimation of β_0 and β_G for Units 3 and 9 were noted due to quasi-complete separation of the data. However, at the end of the 1,000 iteration burn-in the values of the Brooks-Gelman-Rubin statistic were 1.050 and 1.034 for Unit 3, and 1.048 and 1.026 for Unit 9, indicating that all five chains were sampling from the same distribution. Trace and Brooks-Gelman-Rubin statistic plots for Unit 3 for the burn-in and a further 10,000 interaction are shown in Appendix F.4. These suggested that the parameter estimates comprised samples from the whole target distribution.

Using the example of Unit 16, the estimated SMRs varied from 1.22 to 1.46, compared to an estimate of 1.23 from the frequentist analysis. Unsurprisingly, given the amount of data in each case, the estimates for β_{R0} and β_{RG} were more stable than those for β_0 and β_G. The scenario that produced an estimated value for the SMR closest to the frequentist estimate was the final scenario, where the distribution *Normal*(0,1) was specified as the prior distribution for each β_K. This sensitivity of the estimated values of the SMR to the prior distributions of $\boldsymbol{\beta}_K$ was unsurprising as the values for these parameter estimates were, in many cases, derived from a small number of data. A similar pattern of results was seen with all of the other units (details for Units 3 and 8 are given in Appendix F.4).

In the next Section, the results reported are those from the scenario using the distribution *Normal*(0, 1000^2) as the prior distribution for β_0, β_G, β_{R0} and β_{RG} and *Normal*(0, 1) for each parameter β_K. This approach was chosen as it gave a point estimate for the SMR closest to the value from the frequentist analyses.

5.8.5 Results

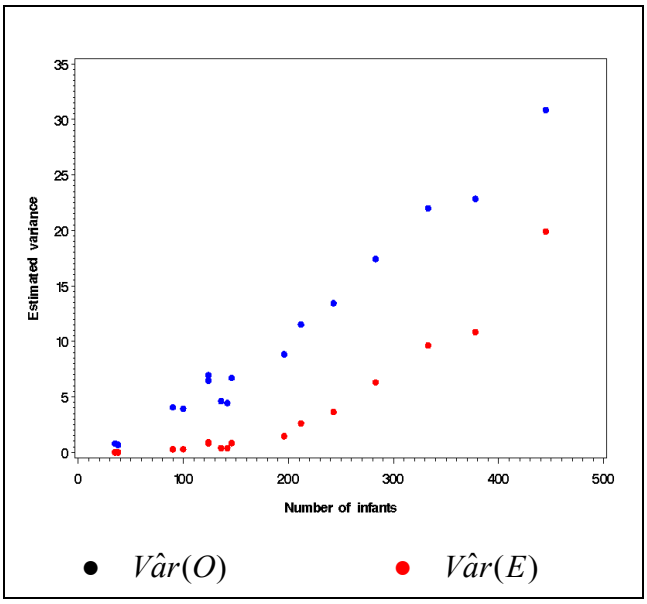
For two of the methods investigated (*Normal approximation* and *BCa bootstrap*) only uncertainty from the observed number of deaths was included. Their use was justified by the observation that the uncertainty associated with the expected number of deaths ($\hat{Var}(E)$) is less than the uncertainty associated with the observed deaths ($\hat{Var}(O)$**).** Table 5.19 shows $\hat{Var}(O)$ and $\hat{Var}(E)$ derived using the methods based on the Normal approximation.

As expected, $\hat{Var}(O)$ is always greater than $\hat{Var}(E)$. However, it can also be seen (perhaps more clearly in Figure 5.17) that the relative difference decreases as the unit size increases. The amount of data in the logistic regression model used to estimate $\hat{Var}(E)$ is reduced when considering large units and so the uncertainty is increased. Hence, ignoring $\hat{Var}(E)$ will result in the estimated confidence intervals being too narrow.

Table 5.19 Estimated variance by NICU

Unit	No. Infants	Observed deaths $\hat{O}$	Expected deaths $\hat{E}$	$\hat{Var}(O)$	$\hat{Var}(E)$
1	212	21	17.8	11.52	2.62
2	283	30	27.9	17.43	6.31
3	38	2	1.1	0.69	0.01
4	142	6	5.3	4.45	0.38
5	333	41	31.1	22.00	9.64
6	378	54	35.1	22.85	10.85
7	243	29	19.7	13.45	3.64
8	124	8	9.8	6.47	0.81
9	35	1	0.9	0.80	0.01
10	146	5	8.9	6.72	0.84
11	445	62	46.9	30.85	19.91
12	196	5	11.5	8.84	1.47
13	136	3	6.0	4.63	0.40
14	90	2	5.2	4.08	0.29
15	124	10	9.9	6.96	0.93
16	100	6	4.9	3.94	0.29

Figure 5.17 Estimated variances by unit size



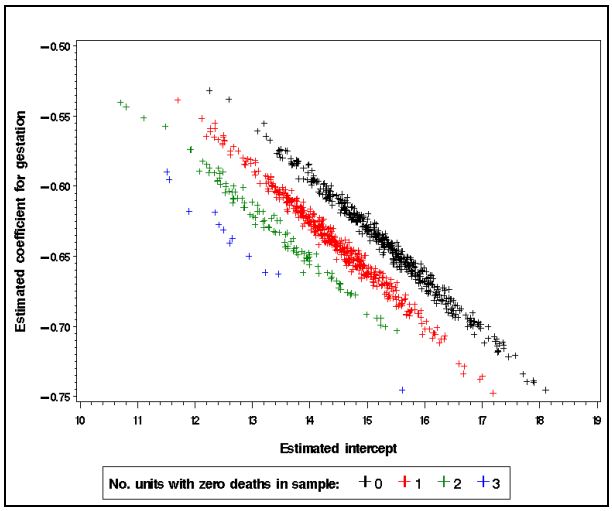
The estimated confidence (credible) intervals are shown in Table 5.20 and Figure 5.19.

Table 5.20 SMR 95% confidence intervals by estimation method

Unit	Ratio	Normal approx.	BCa Bootstrap	Hosmer & Lemeshow	Zhou & Romano	Bootstrap	Bayesian	
1	1.18	0.80 to 1.56	0.83 to 1.58	0.82 to 1.71	0.76 to 1.60	0.87 to 4.97	1.20	0.85 to 1.65
2	1.07	0.78 to 1.37	0.77 to 1.45	0.77 to 1.49	0.73 to 1.42	0.80 to 4.49	1.09	0.75 to 1.52
3	1.80	0.33 to 3.27	0.00 to 6.23	0.78 to 4.13	0.32 to 3.28	0.00 to 12.91	1.85	0.94 to 3.14
4	1.12	0.34 to 1.90	0.47 to 2.03	0.54 to 2.32	0.31 to 1.93	0.53 to 6.60	1.12	0.49 to 2.10
5	1.32	1.02 to 1.62	1.02 to 1.65	0.97 to 1.78	0.96 to 1.68	1.05 to 6.04	1.34	0.98 to 1.81
6	1.54	1.27 to 1.81	1.25 to 1.88	1.19 to 1.99	1.21 to 1.87	1.23 to 5.95	1.57	1.19 to 2.05
7	1.47	1.10 to 1.84	1.07 to 1.96	1.07 to 2.01	1.05 to 1.88	1.06 to 6.72	1.49	1.05 to 2.06
8	0.82	0.30 to 1.34	0.40 to 1.34	0.42 to 1.58	0.27 to 1.37	0.42 to 4.16	0.82	0.40 to 1.43
9	1.09	-0.83 to 3.02	0.00 to 2.98	0.18 to 6.45	-0.84 to 3.04	0.00 to 5.09	1.11	0.05 to 2.40
10	0.56	0.00 to 1.14	0.19 to 1.18	0.19 to 1.59	-0.04 to 1.17	0.20 to 3.85	0.55	0.20 to 1.13
11	1.32	1.08 to 1.56	1.07 to 1.60	1.02 to 1.71	1.02 to 1.62	1.06 to 5.22	1.35	1.04 to 1.74
12	0.44	-0.07 to 0.95	0.14 to 0.92	0.13 to 1.43	-0.11 to 0.99	0.13 to 2.74	0.42	0.14 to 0.91
13	0.50	-0.20 to 1.21	0.11 to 1.17	0.12 to 2.07	-0.23 to 1.24	0.00 to 2.79	0.47	0.13 to 1.14
14	0.38	-0.37 to 1.14	0.00 to 1.19	0.05 to 2.79	-0.39 to 1.17	0.00 to 2.24	0.35	0.05 to 1.07
15	1.01	0.48 to 1.54	0.57 to 1.63	0.58 to 1.77	0.45 to 1.58	0.56 to 5.08	1.02	0.57 to 1.63
16	1.23	0.43 to 2.03	0.49 to 2.35	0.61 to 2.44	0.40 to 2.06	0.50 to 7.02	1.22	0.50 to 2.36

The confidence intervals obtained using the second (percentile) bootstrap method are very wide. This arose because some bootstrap samples had units with no sampled deaths. This caused the problem of data separation in the model (§3.4) and resulted in poorly estimated model parameters with large standard errors. Figure 5.18 shows the grouping of the estimated values of β_0 and β_G according to the number of reference units with no sampled observed deaths.

Figure 5.18 Estimated coefficients by number of units with sampled zero deaths



The details of this behaviour will not be investigated in this thesis. It is sufficient to note that this method is difficult to apply when some of the units are small, or where the outcome is rare. The problem may be overcome by using a different model parameterization, for example the 'rest of Region' parameterization in §5.3.1. The estimated SMRs from such a model are shown in Table 5.21. However, the different parameterisation of the units means that any estimates from this approach are not directly comparable to the estimates from the deviation contrast models explored elsewhere in this Section.

The differences between the methods for the deviation parameterisation model (Table 5.20) are most marked for small units; e.g. Units 3 & 9. As was seen in the simulation study, the estimates obtained using the method of Hosmer & Lemeshow tended to yield higher values for upper limit, and negative lower limits were observed for the Normal approximation and Zhou & Romano method. The differences in the estimated limits from the normal approximation and Zhou & Romano methods were small: the addition of the uncertainty in the total number of expected deaths had little effect compared to the difference between methods. The point estimates from the Bayesian methods differed from the MLE point estimates due to the influence of the prior distributions.

Table 5.21 SMR from 'Rest of Region' model, with bootstrapped 95% confidence intervals

Unit	SMR	(95% CI)
1	1.00	(0.67 to 1.31)
2	0.92	(0.65 to 1.25)
3	1.52	(0.00 to 4.27)
4	0.93	(0.34 to 1.57)
5	1.11	(0.84 to 1.41)
6	1.35	(1.07 to 1.70)
7	1.25	(0.86 to 1.66)
8	0.70	(0.30 to 1.1)
9	0.90	(0.00 to 2.31)
10	0.48	(0.12 to 0.88)
11	1.15	(0.89 to 1.40)
12	0.37	(0.08 to 0.76)
13	0.43	(0.00 to 0.91)
14	0.33	(0.00 to 0.93)
15	0.86	(0.43 to 1.32)
16	1.01	(0.34 to 1.85)

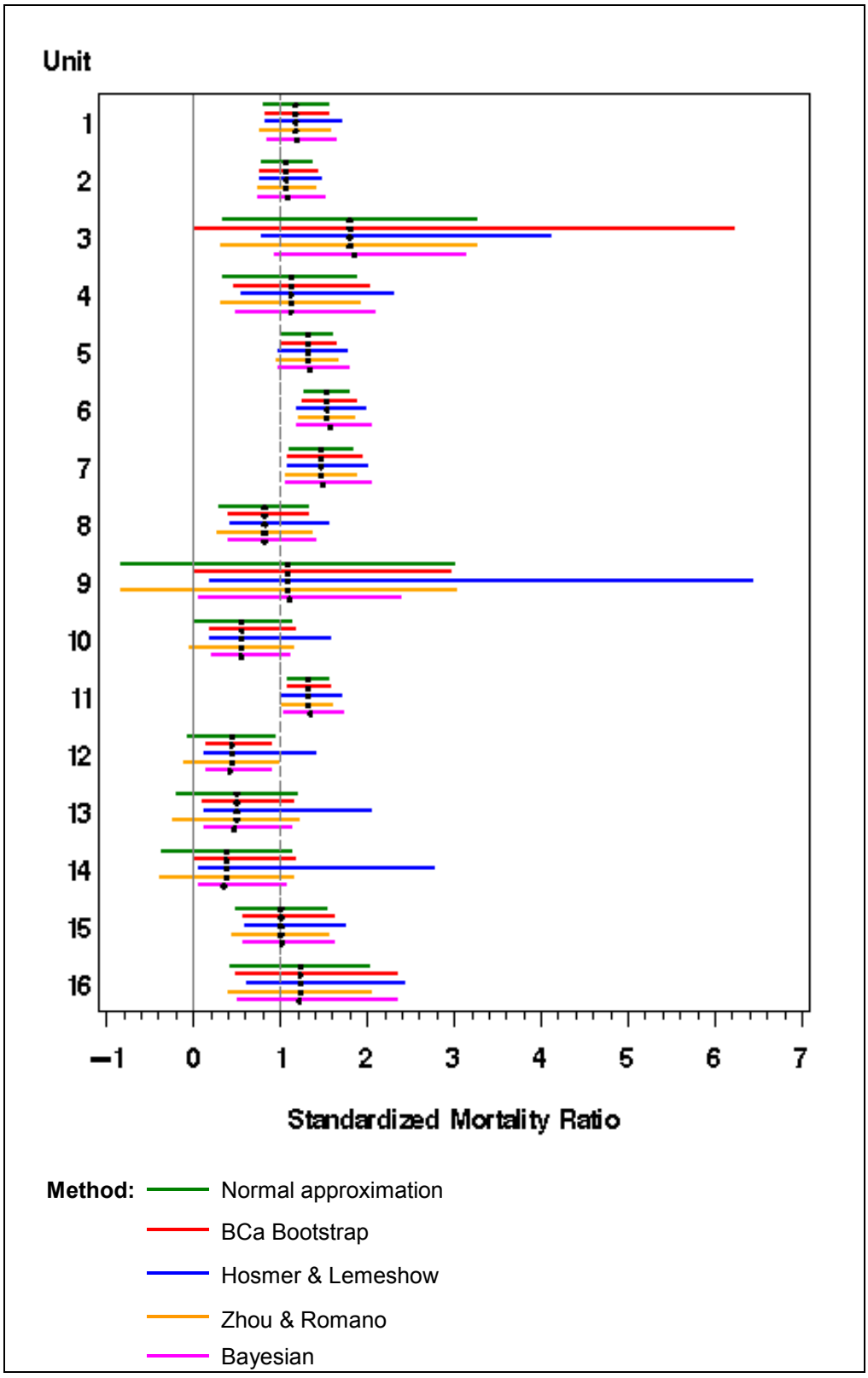
5.8.6 Choice of confidence interval estimation method

The method proposed by Hosmer and Lemeshow (1995) and the second bootstrap method offered advantages over the other frequentist approaches: (i) they included all uncertainty (although comparison of the intervals estimated using the Normal approximation and those using the method proposed by Zhou and Romano (1997) suggest that the effect of including $\hat{Var}(E)$ is not great); (ii) they are symmetrical on the logarithmic scale; (iii) the simulation study shows evidence of good coverage properties.

The bootstrap also has one other useful characteristic over the methods proposed by Hosmer & Lemeshow and Zhou & Romano. As the estimate of uncertainty only depends on the estimated probability of mortality, and not the estimated variances of the parameter estimates, it is not influenced by collinearity. A major problem with collinearity would inflate the variance of the parameter estimates and, therefore, inflate the width of the Hosmer & Lemeshow intervals, since they are estimated using the covariance matrix $\mathbf{S}\left(\hat{\boldsymbol{\beta}}_{R(-j)}\right)$ (*5.47*).

The point estimates used in the bootstrap method are unaffected by collinearity (Harrell, 2001:64-65).

Figure 5.19 Confidence and credible intervals for SMRs in TNS data, using different estimation methods



If the problem of small data sets and the sampling of zero deaths did not occur in these data, the 'full' bootstrap method would, perhaps, be the most appropriate choice. It has all the desirable properties of the method proposed by Hosmer & Lemeshow, except that it is computationally intensive. However, as was shown above, this option is not suitable for these data, due to the small unit sizes, and the method of Hosmer & Lemeshow was preferred.

The Bayesian method also showed good properties. This approach will also be illustrated in Chapter 6.

5.9 Mortality Difference

The SMR and its use have been discussed in the previous Section. However, an alternative method of presentation is to report the difference between the number of observed and the number of expected deaths:

$$\text{Mortality difference for Unit j} = Observed_j - Expected_j = \sum_{i=1}^{n_j} d_i - \sum_{i=1}^{n_j} \pi_i \qquad (5.57)$$

If, applying indirect standardization, it is assumed that the fitted logistic model is correct then the expected number of deaths is estimated by $\sum_{i=1}^{n} \hat{\pi}_i$.

Therefore, the difference between the observed and expected mortality, ***w***, is estimated by:

$$\hat{w} = \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} \hat{\pi}_i$$

The mortality difference, "*excess mortality*", has been used for provider profiling, for example in the report of the Bristol Royal Infirmary Inquiry (Spiegelhalter *et al,* 2002). Confidence intervals for such differences can be estimated using methods similar to those for the SMR.

The standardized mortality difference (*w*) is illustrated using the TNS data (Table 5.22). The 95% confidence intervals were estimated using the Normal approximation method, without the inclusion of the continuity correction (Glance *et al,* 2000):

$$\left(\sum_{i=1}^{n_j} d_i - \sum_{i=1}^{n_j} \hat{\pi}_i \right) \pm z_{1-\alpha/2} \sqrt{\sum_{j=1}^{N} \left(\hat{\pi}_i \left[1 - \hat{\pi}_i \right] \right)}$$

where: $d = \begin{cases} 1 & \text{if died before discharge} \\ 0 & \text{if discharged alive} \end{cases}$

$\hat{\pi}$ is the estimated probability of death before discharge

The absolute value of any difference needs to be interpreted in the light of the workload of the unit. The difference between the observed and expected mortality does not allow for the relative size of a particular unit. This is important when interpreting such results. The following illustration is given in Hosmer and Lemeshow (1995):

> *"For example, suppose that among 100 patients the observed number of deaths in a particular ICU is 10 and the expected number is 7. In this case, the difference is 3 and the ratio is 1.43. In a second ICU, assume that among 100 patients the observed number of deaths is 30 while the expected number is 27. Again, the difference between observed and expected is 3 but now the ratio is 1.11. From the point of view of interpretability, the ratio provides evidence that hospital B performs at a higher level than hospital A (11 per cent excess mortality versus 43 per cent) whereas, based on the difference measure, there is no difference between the ICUs."*

Table 5.22 Difference between observed and expected mortality

Unit	Observed Deaths (O)	Expected Deaths (E)	*w* (O-E)	Var(O)	(95% CI)	P(O=E)
1	21	17.9	3.1	11.40	(-3.5 to 9.8)	0.36
2	30	27.6	2.4	17.65	(-5.8 to 10.7)	0.56
3	2	1.1	0.9	0.69	(-0.7 to 2.6)	0.28
4	6	5.3	0.7	4.43	(-3.4 to 4.8)	0.75
5	41	31.3	9.7	21.91	(0.5 to 18.9)	0.039
6	54	34.8	19.2	22.95	(9.7 to 28.6)	0.0001
7	29	19.7	9.3	13.45	(2.0 to 16.5)	0.011
8	8	9.7	-1.7	6.48	(-6.7 to 3.3)	0.49
9	1	0.9	0.1	0.80	(-1.6 to 1.9)	0.92
10	5	8.9	-3.9	6.72	(-8.9 to 1.2)	0.13
11	62	46.9	15.1	30.87	(4.2 to 26.1)	0.0064
12	5	11.5	-6.5	8.88	(-12.3 to -0.7)	0.030
13	3	6.0	-3.0	4.62	(-7.2 to 1.3)	0.16
14	2	5.2	-3.2	4.08	(-7.2 to 0.8)	0.11
15	10	9.9	0.1	6.94	(-5.0 to 5.3)	0.96
16	6	4.9	1.1	3.95	(-2.7 to 5.0)	0.58

To overcome this problem a standardized difference has been proposed: the *W*-statistic (Sacco *et al*, 1994):

$$W = \left(\frac{\sum_{i=1}^{n_j} d_i - \sum_{i=1}^{n_j} \pi_i}{n} \right) \times 100$$

Hence, the value of the W-statistic represents the number of 'excess deaths' per 100 patients, thus allowing interpretation of the statistic to be independent of the number of observations. However, whichever way the difference is reported (*w* or *W*) the same difficulties arise in comparing values across units as was discussed for the SMR (§5.5.2) since indirect standardization was used to obtain the expected number of deaths, (Glance *et al*, 2000). To overcome this, a standardized *W*-statistic W_S has been proposed (Hollis *et al*, 1995), analogous to the standardized SMR discussed in §5.5.3 and with the same difficulties in application.

In general, the mortality ratio (O/E) is to be preferred to a statistic based on mortality difference as a ratio, rather than the difference, is likely to allow a better interpretation of the effect size to be made for each unit. Therefore, the SMR is used in Chapter 6.

5.10 Random Effects Modelling

The data used in this thesis have a hierarchical structure, with infants nested within hospitals. It has been argued that in such situations statistical methods that take the data structures into account should be used (Goldstein and Spiegelhalter, 1996; Normand *et al,* 1997). It is suggested that the possible correlation between observations within the same cluster should be accounted for, otherwise the estimates of the standard errors may be biased (Goldstein, 1995:3). The assumption is made that, after adjustment for individual-level and NICU-level covariates, the units are exchangeable, that is, they can be seen as having been drawn from a population of units with a specified probability distribution. Such methods produce shrunken estimates of the unit effects, with the effects shrunk towards the population mean. The model can be expressed, assuming in this case that the unit effects follow a Normal distribution (**the mixing distribution**), as:

$$g_i = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \delta_j \qquad (5.58)$$

where: $\delta \sim Normal(0, \sigma^2)$

The values for the parameters β_0, $\boldsymbol{\beta}$ and σ are estimated from the data.

The values δ_j represent the units' effects and can be used as a measure of a unit's deviation from the Regional average, for example Leyland and Boddy (1998), Goldstein and Spiegelhalter (1996), DeLong *et al* (1997), Normand *et al* (1997).

5.10.1 Mixing distribution

If a random effects model is to be used, there are difficulties in applying such an approach. It is possible, especially if an important unit level covariate is missing, that the random effects do not follow a Normal distribution (Marshall and Spiegelhalter, 1998a; Mohammed *et al,* 2001b). It is particularly unclear in non-linear modelling whether a Normal mixing distribution is appropriate. Previous research has shown that misspecification of the mixture distribution of a logistic regression model introduces little bias into the estimates of the fixed parameters; i.e. $\boldsymbol{\beta}$ (Agresti *et al*, 2004; Butler and Louis, 1992; Neuhaus *et al*, 1992). However, under such misspecification, inferences concerning the mixing distribution are less robust (Aylin *et al,* 2001b; Neuhaus *et al*, 1992; Turner *et al*, 2001; Verbeke and Lesaffre. 1996). This, obviously, has important implications for the estimation of δ_j.

Non-Normal mixing distributions

Although methods exist to try to assess the assumption of Normality of the random effects for linear mixed models (Jaing, 2001; Lange and Ryan, 1989; Ritz, 2004) these are not straightforward and their application to non-linear mixed models undeveloped. In addition, statistical software that allows the specification of a non-Normal mixing distribution within a frequentist framework is not generally available. Alternative distributions should be possible in SAS `PROC NLMIXED` but, as of version 8.2, this feature has not been introduced. The only option available to this thesis was the `gnlmix` function within the `REPEATED` package for the *R* software (Lindsey, 2005). This function, using Romberg integration, allows different mixing distributions to be specified. However, the estimation of level 2 residuals is not available. Nevertheless, this package does allow an investigation into the appropriateness of the assumption of a Normal mixing distribution.

In addition to the Normal distribution, the three alternative probability distributions were investigated: Cauchy, Laplace (double exponential) and Logistic. These distributions were chosen as they can take values over the whole real line. To avoid overparameterisation, in each case the mean (mode in the case of the Cauchy distribution) is set to zero. It was noted that these distribution are all symmetrical about the mean, which may not reflect the true distribution of the units. However, available alternative non-symetrical distributions, for example the Gamma and Beta distributions, are bounded and, therefore, unsuitable for these data.

5.10.2 Application to TNS data

A model was analysed using gestational age as the only risk-adjustment variable (in keeping with the rest of this chapter):

$$g_i = \beta_0 + \beta_G . gest_i + \delta_j \qquad (5.59)$$

where: *gest* is the gestational age at birth, in completed weeks

Four models were considered, each with a different probability distribution specified for δ: Normal, Cauchy, Laplace and Logistic.

As suggested previously, the estimates for the fixed parameters are reasonably robust to the choice of mixture distribution (Table 5.23). In this case, the variance estimates for the random effects were also similar.

Table 5.23 Parameter estimates from random effects models using various mixture distributions

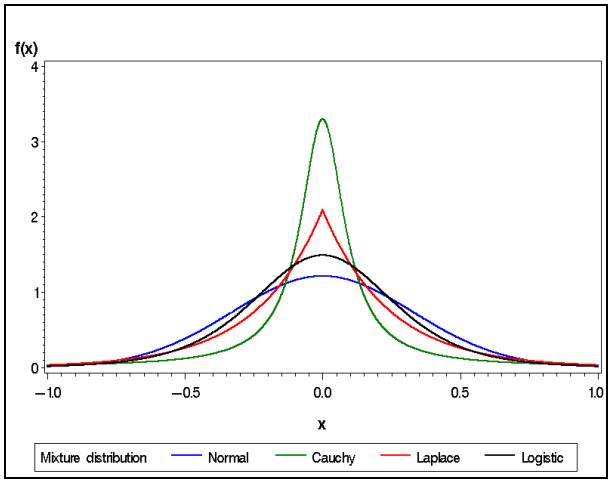
Distribution	$\hat{\beta}_0$	**(s.e.)**	$\hat{\beta}_G$	**(s.e.)**	$\hat{\sigma}^2$	**AIC**
Normal	15.8521	(0.9015)	-0.6566	0.0335	0.1083	634.6552
Cauchy	16.0111	(0.8909)	-0.6585	0.0335	- [j]	635.9348
Laplace	15.9219	(0.8954)	-0.6575	0.0335	0.1141	634.9994
Logistic	15.8800	(0.8994)	-0.6570	0.0335	0.0926	634.7929

The four estimated mixture distributions are shown in Figure 5.20. The estimated Cauchy distribution was the most different of the four, being more a leptokurtic distribution. However, the value for the Akaike Information Criterion (AIC) was lowest for the Normal

[j] Variance does not exist for the Cauchy distribution (Rothschild and Logothetis, 1986)

distribution (albeit by a very small amount), therefore offering no evidence that any of the alternative distributions offered an improvement in the model (Table 5.23).

Figure 5.20 Estimated mixture distributions



Estimation of level-2 residuals

Since the `glnmix` function in *R* does not facilitate the estimation of the value of the residuals for the random effects (Lindsey, J.: Personal communication), the model was re-estimated with a Normal mixture distribution in SAS `PROC NLMIXED`, using the default dual quasi-Newton algorithm (SAS Institute Inc., 1999:2460-2465). The parameter estimates obtained were very similar to those from *R*'s `gnlmix` (Table 5.23), differing only because of the different estimation procedures in each package:

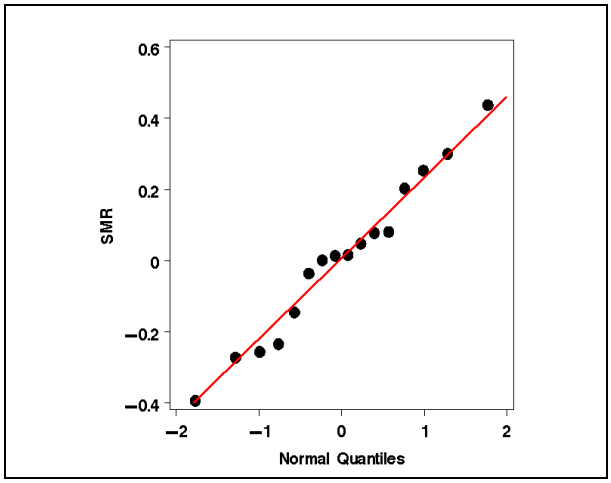
$\hat{\beta}_0 = 15.8520$ (s.e. = 0.9040)

$\hat{\beta}_G = -0.6566$ (s.e. = 0.0335)

$\hat{\sigma}^2 = 0.1081$ (s.e. = 0.0866)

`PROC NLMIXED` allowed the estimation of unit level residuals (Table 5.24). Although it is difficult to determine any pattern with only 16 observations, a q-q plot showed evidence that these residuals do seem to follow a Normal distribution (Figure 5.21).

Figure 5.21 Q-Q plot for level 2 residuals



The residuals can be presented together with estimated 95% confidence intervals. More informatively, by taking the exponential the residuals can be presented as estimated odds ratios, comparing each unit to the mean of the mixture distribution (Table 5.24). In all cases the estimated 95% confidence interval contains the value unity, offering no statistical evidence that any unit is different from the Regional 'average'.

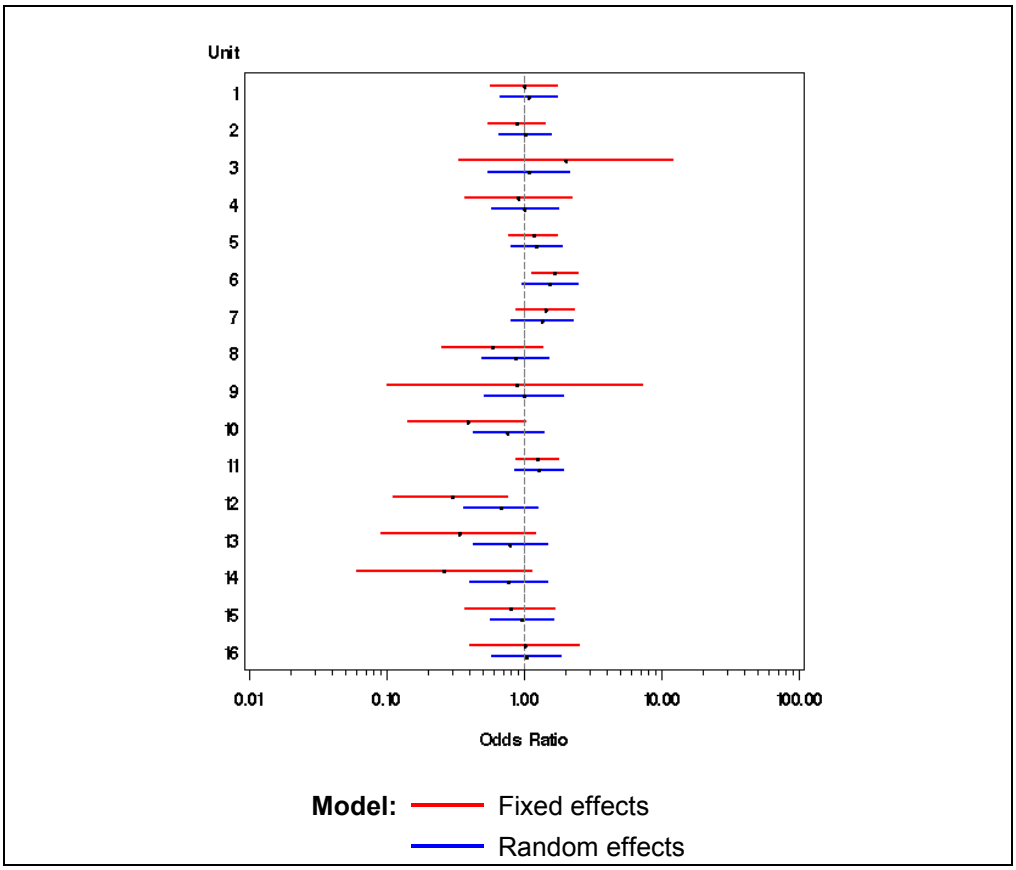
Table 5.24 Estimated unit level residuals and odds ratios from mixed model

Unit	$\hat{\delta}_j$	(s.e.)	(95% C.I.)	$\hat{\omega}_j$	(95% C.I.)
1	0.081	(0.234)	(-0.416 to 0.579)	1.08	(0.66 to 1.78)
2	0.016	(0.211)	(-0.433 to 0.465)	1.02	(0.65 to 1.59)
3	0.078	(0.322)	(-0.608 to 0.765)	1.08	(0.54 to 2.15)
4	0.014	(0.271)	(-0.565 to 0.593)	1.01	(0.57 to 1.81)
5	0.203	(0.209)	(-0.242 to 0.648)	1.23	(0.79 to 1.91)
6	0.437	(0.224)	(-0.042 to 0.915)	1.55	(0.96 to 2.50)
7	0.300	(0.245)	(-0.222 to 0.823)	1.35	(0.80 to 2.28)
8	-0.145	(0.263)	(-0.706 to 0.416)	0.87	(0.49 to 1.52)
9	0.001	(0.315)	(-0.670 to 0.672)	1.00	(0.51 to 1.96)
10	-0.272	(0.285)	(-0.879 to 0.336)	0.76	(0.42 to 1.40)
11	0.253	(0.197)	(-0.166 to 0.673)	1.29	(0.85 to 1.96)
12	-0.393	(0.298)	(-1.028 to 0.243)	0.68	(0.36 to 1.28)
13	-0.234	(0.300)	(-0.873 to 0.406)	0.79	(0.42 to 1.50)
14	-0.256	(0.312)	(-0.922 to 0.410)	0.77	(0.40 to 1.51)
15	-0.036	(0.252)	(-0.574 to 0.503)	0.96	(0.56 to 1.65)
16	0.048	(0.277)	(-0.543 to 0.639)	1.05	(0.58 to 1.89)

5.10.3 Comparison with estimates from fixed-effects model

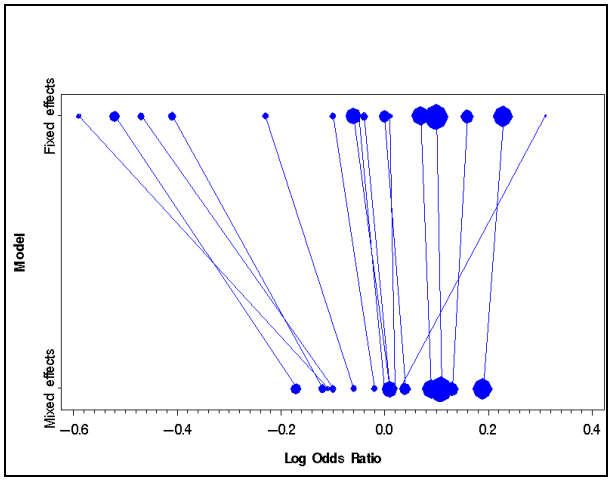
The estimates from Table 5.24 were compared to the results from a fixed-effects model, in this case the 'Rest of Region' model (Table 5.1) although all potential parameterisations of the units showed the same pattern. The point estimates from the random-effects model were obviously 'shrunken' towards the mean value (Figure 5.22). The widths of the confidence intervals from the random-effects model were also narrower than those from the fixed-effects model.

Figure 5.22 Estimated Odds Ratios from fixed-effects and random-effects models, with 95% confidence intervals



The point estimates are also compared in Figure 5.23. In this Figure the areas of the circles are proportional to the number of admissions to the neonatal units. Although the values zero are not directly comparable, they represent different Regional 'averages', the shrinkage induced into the estimates with the random-effects model, particularly for the smaller unit, can once again be clearly be seen, with the estimates for the smaller units changing the most.

Figure 5.23 Comparison of log odds ratios from fixed-effects and random-effects models



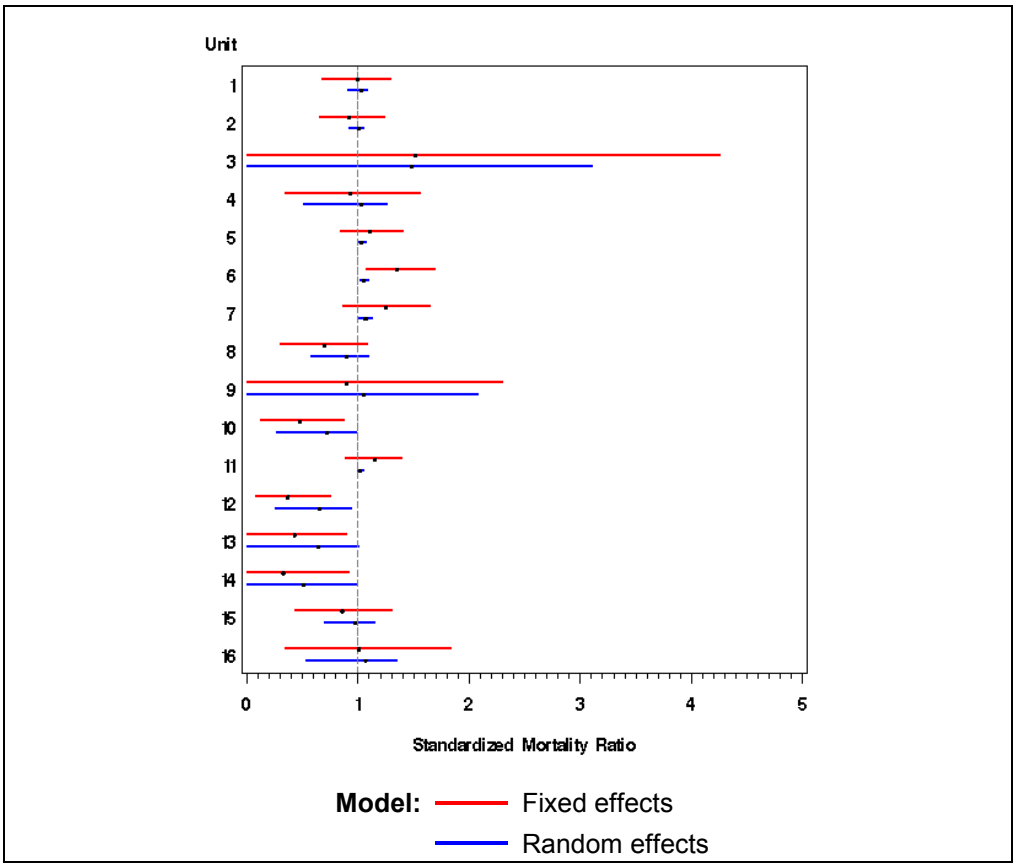
The results can also be presented as Standardized Mortality Ratios (SMRs). SAS `PROC NLMIXED` was used to estimate a SMR for each unit using the model shown in equation (*5.58*). Ninety-five percent confidence intervals were estimated using the percentile bootstrap method described in §5.6.3 that allowed for uncertainty in both the observed and expected number of deaths.

Table 5.25 Estimated SMR, with 95% confidence interval, from mixed model

Unit	SMR	(95% CI)
1	1.03	(0.91 to 1.10)
2	1.01	(0.92 to 1.06)
3	1.48	(0.00 to 3.12)
4	1.03	(0.51 to 1.27)
5	1.03	(0.99 to 1.08)
6	1.05	(1.02 to 1.11)
7	1.07	(1.00 to 1.14)
8	0.90	(0.57 to 1.11)
9	1.05	(0.00 to 2.09)
10	0.72	(0.27 to 1.00)
11	1.02	(1.00 to 1.06)
12	0.66	(0.25 to 0.95)
13	0.64	(0.00 to 1.02)
14	0.51	(0.00 to 1.00)
15	0.98	(0.70 to 1.16)
16	1.07	(0.53 to 1.36)

The estimated SMRs (Table 5.25) took values closer to the value one than the equivalent estimates from the fixed effects model (Table 5.21): i.e. they were 'shrunken' estimates. In addition, the estimated 95% confidence intervals were generally very narrow, especially for the large units (Figure 5.24). Although the estimates were not precisely comparable, as all 16 units were used to provide the estimate of expected number of deaths for the mixed model, whereas for the fixed-effects model the observations from the unit of interest were excluded from the model, the models were sufficiently similar to make the comparison of interest.

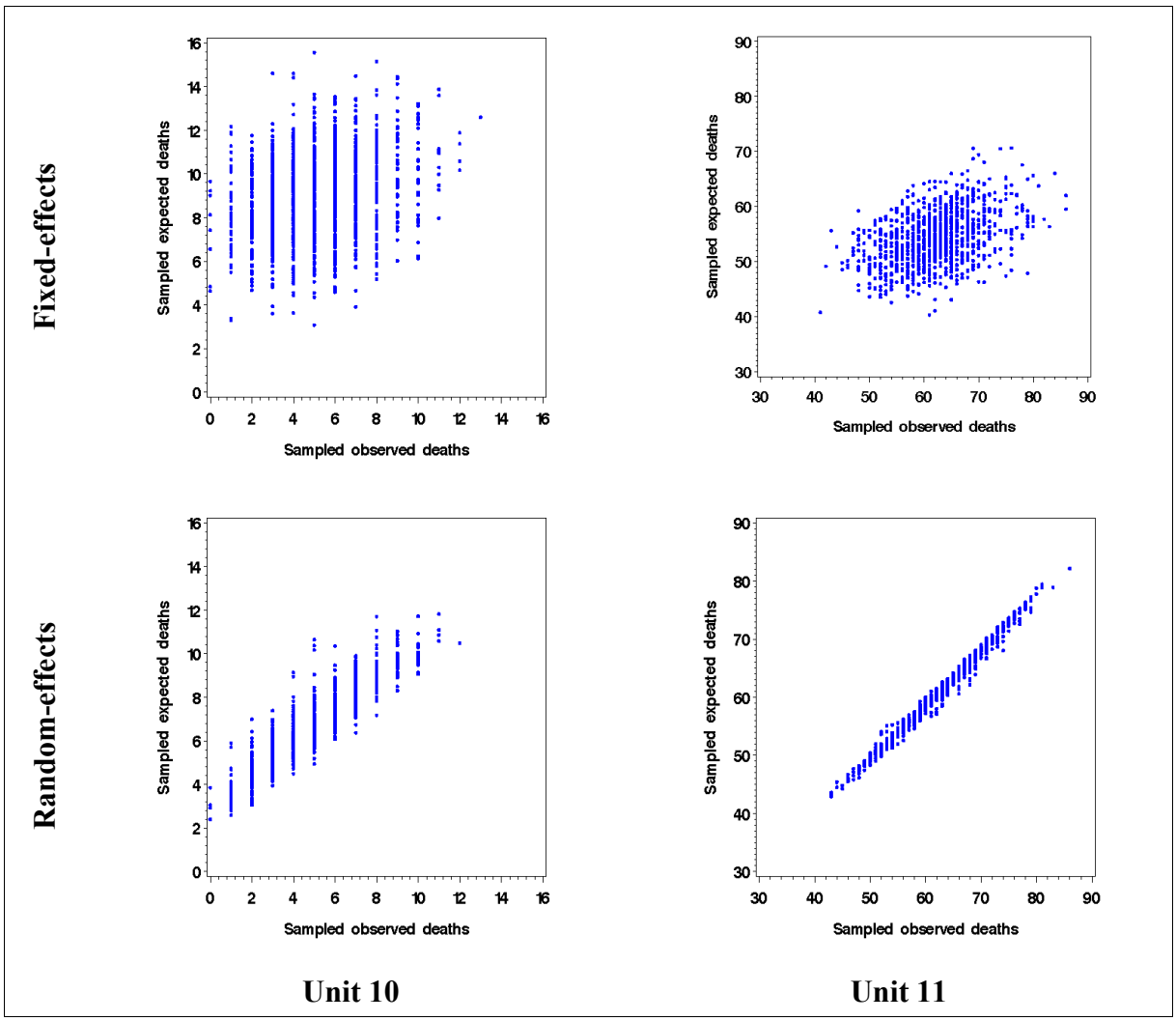
Figure 5.24 Estimated SMRs from fixed-effect and random-effect models



The narrow confidence intervals from the random-effects model are the result of obtaining, for each bootstrap sample, an estimated expected number of deaths close to the sampled observed number. Examples of this are shown for Units 10 and 11 (Figure 5.25).

The exact cause of this phenomenon was not investigated but was likely to have been due to shrunken estimates of the unit effects being estimated in the random-effects model.

Figure 5.25 Estimated expected number of deaths and sampled number of death for fixed-effects and random-effects models



5.10.4 Intraclass correlation

Exact methods for estimating the **intraclass correlation** (ICC) do not exist for hierarchical logistic regression models. The approach used for linear hierarchical models does not apply here as the variance and the mean are linked, meaning that the value of the ICC is dependent of the rate of positive outcome in the sample, and the level 2 variance is measured on the logistic scale. Different approximations have been proposed and it has been suggested that these lead to similar results (Chaix et al, 2004). A straightforward approximate estimator (ρ_I) has been put forward by Snijders & Bosker (1999):

$$\rho_{\mathrm{I}} = \frac{\sigma^2}{\sigma^2 + \pi^2/3}$$

Substituting $\hat{\sigma}^2 = 0.1081$ gives:

$$\hat{\rho}_{\mathrm{I}} = 0.0318$$

Such a low value for the ICC suggests small differences in performance between the units.

5.10.5 The Bayesian approach

The Bayesian approach also allows the introduction of alternative distributions and can be implemented in a straightforward manner using standard software (e.g. WinBUGS). However, it has been shown that parameter estimates are sensitive to the choice of prior distribution, together with its variance, for the random effects (Turner *et al*, 2001).

The model was briefly explored here by specifying a range of prior probability distributions for the mixture distribution. The results were inspected by reported the observed point estimates for $\hat{\delta}_j : j \in \{1, 2, \ldots, 16\}$. This approach was taken, rather than reporting the estimated SMRs, as interest was on the random-effects model and not on the estimation of the uncertainty in the observed number of deaths (as was the case in §5.8).

The model can be specified, from (*5.59*), as:

$$g = \beta_0 + \beta_G . gest_i + \delta_j \qquad j \in \{1, 2, \ldots, 16\}$$

Prior probability distributions were required for the model parameters. For the parameters β_0 and β_{G} the distribution $Normal(0, 1000^2)$ was used. Three distributions were specified for the random effects δ: Normal (*N*), Logistic (*L*) and Laplace or Double Exponential (*DE*), each with mean zero and variance σ^2. Three prior distribution given to σ^2 in the case of the Normal distribution: $1/\sigma^2 \sim Gamma(0.0001,0.0001)$, $\sigma^2 \sim Uniform(0,1000)$ and $\sigma^2 \sim Uniform(0,10)$. For the Logistic and Laplace distributions $\sigma^2 \sim Uniform(0,1000)$ was used.

For each model five chains were inspected over a 1,000 interation burn-in. The first specification of the model produced poor mixing (Brooks-Gelman-Rubin statistic plots are shown in Appendix E.3):

$$g = \beta_0 + \beta_G . gest_i + \delta_j \qquad j \in \{1, 2, \ldots, 16\}$$

with prior probability distributions:

$$\beta_0 \sim N(0,1000^2)$$
$$\beta_G \sim N(0,1000^2)$$
$$\delta \sim N(0,\sigma^2)$$
$$\sigma^2 \sim U(0,1000)$$

This problem was solved by specifying the (mathematically equivalent) model:

$$g = \beta_G.gest_i + \delta_j \qquad j \in \{1, 2, \ldots, 16\}$$

with prior probability distributions:

$$\beta_0 \sim N(0,1000^2)$$
$$\beta_G \sim N(0,1000^2)$$
$$\delta \sim N(\beta_0,\sigma^2)$$
$$\sigma^2 \sim U(0,1000)$$

This method of model specification model showed good sampling properties for all scenarios (Appendix E.3) and a further 10,000 iteration were used in each case to estimate values for the model parameters (Table 5.26).

The choice of prior distribution for the random effect had little influence on the estimated values of the fixed model parameters $\hat{\beta}_0$ and $\hat{\beta}_G$. It was seen in §5.10.2 that, within a frequentist framework, the estimates of the fixed parameters were robust to the choice of mixture distribution and there is evidence here that this also applies to the Bayesian estimates.

However, the values of the parameter estimates for the random part of the model are influenced both by the choice of probability distribution and by the prior distribution for the variance. If such an approach was to be used in practice very careful consideration would need to be given to the choice of prior distributions.

Table 5.26 Estimates from Bayesian random-effects model

Estimate	Model					
	$\beta_0 \sim N(0,1000^2)$ $\beta_G \sim N(0,1000^2)$ $\delta \sim N(\beta_0, \sigma^2)$ $\sigma^2 \sim U(0,1000)$	$\beta_0 \sim N(0,1000^2)$ $\beta_G \sim N(0,1000^2)$ $\delta \sim N(\beta_0, \sigma^2)$ $1/\sigma^2 \sim G(0.0001,0.0001)$	$\beta_0 \sim N(0,1000^2)$ $\beta_G \sim N(0,1000^2)$ $\delta \sim N(\beta_0, \sigma^2)$ $\sigma^2 \sim U(0,10)$	$\beta_0 \sim N(0,1000^2)$ $\beta_G \sim N(0,1000^2)$ $\delta \sim L(\beta_0, \tau)$ $Var(\delta) \sim U(0,1000)$	$\beta_0 \sim N(0,1000^2)$ $\beta_G \sim N(0,1000^2)$ $\delta \sim DE(\beta_0, \lambda)$ $Var(\delta) \sim U(0,1000)$	Frequentist estimates [k]
$\hat{\delta}_1$	0.113	0.061	0.113	0.096	0.051	0.081
$\hat{\delta}_2$	0.043	0.012	0.043	0.027	-0.001	0.016
$\hat{\delta}_3$	0.129	0.060	0.125	0.115	0.080	0.078
$\hat{\delta}_4$	0.025	0.007	0.023	0.018	-0.001	0.014
$\hat{\delta}_5$	0.249	0.172	0.249	0.223	0.161	0.203
$\hat{\delta}_6$	0.501	0.395	0.499	0.483	0.444	0.437
$\hat{\delta}_7$	0.367	0.258	0.364	0.338	0.281	0.300
$\hat{\delta}_8$	-0.167	-0.116	-0.175	-0.169	-0.146	-0.145
$\hat{\delta}_9$	-0.002	0.002	0.004	-0.003	-0.004	0.001
$\hat{\delta}_{10}$	0.345	-0.233	-0.348	-0.346	-0.342	-0.272
$\hat{\delta}_{11}$	0.298	0.220	0.298	0.270	0.217	0.253
$\hat{\delta}_{12}$	-0.499	-0.348	-0.489	-0.519	-0.567	-0.393
$\hat{\delta}_{13}$	-0.321	-0.198	-0.316	-0.316	-0.304	-0.234
$\hat{\delta}_{14}$	-0.363	-0.222	-0.366	-0.360	-0.366	-0.256
$\hat{\delta}_{15}$	-0.030	-0.029	-0.032	-0.036	-0.037	-0.036
$\hat{\delta}_{16}$	0.073	0.032	0.073	0.058	0.027	0.048
$\hat{\sigma}^2$	0.206	0.106	0.202	0.224	0.276	0.108
$\hat{\beta}_0$	-3.89	-3.84	-3.89	-3.86	-3.83	-3.85
$\hat{\beta}_G$	-0.66	-0.66	-0.66	-0.66	-0.66	-0.66

5.10.6 Choice between fixed and random effects models

It has previously been shown that random-effects models have greater specificity (true negative rate) than fixed-effects models considered, and that fixed-effects models have greater

[k] The estimated value for β_0 differs from Table 5.24 as gestational age centred at 30 weeks was used here

sensitivity (true positive rate) (Goldstein and Spiegelhalter, 1996; DeLong *et al,* 1997; Austin *et al,* 2003). This has also been suggested in this thesis for both the estimated odds ratio (Figure 5.22) and standardized mortality rate (Figure 5.24). In each case, units were identified as statistical outliers from fixed-effects models but none were detected with random-effects models.

The choice of model will reflect the aims of the investigators (Draper and Gittoes, 2004). A fixed-effects model is more likely to identify true extreme performers, but at the expense of an increased false positive rate. Conversely, a random-effects model will tend to have a higher false negative rate, meaning that some outlying providers are not identified. The aim of the analysis of the data from TNS is to provide a screening of the data, to identify units in which a further (clinical) investigation of mortality rates may be appropriate. It was felt, therefore, that a fixed-effects model was the more appropriate for these data.

5.10.7 Further alternatives

An alternative approach, used for the Bristol Royal Infirmary Inquiry (Spiegelhalter *et al,* 2002), was to use a random-effects model to estimate an 'average' Regional effect using the data without the unit of interest. These model parameters could then be used to produce an indirectly standardized measure of performance. The expected number of deaths for the unit of interest is calculated using the mean value of the distribution of the reference units: i.e. β_0 (from *5.59*). This is an alternative reweighing of the reference units, analogous to the methods discussed in §5.3.1.

In the case of the Bristol Royal Infirmary Inquiry, the difference between the observed and expected mortality was presented but to illustrate the method here the standardized mortality ratio was calculated (Table 5.27). The 95% confidence intervals were estimated by fitting models to 1,000 bootstrap samples and finding the 2.5^{th} and 97.5^{th} percentiles of values of the SMRs from the samples.

The estimates obtained were lower that those from the fixed-effects model with deviation contrasts for the units (Table 5.20) but were greater than the fixed-effects model with the 'Rest of region' parameterisation (Table 5.21). This duplicated the phenomenon seen when estimating odds ratios in §5.3.1, as the use of random-effects models in this way represents a weighted analysis of the units. However, once again the estimated confidence intervals are much narrower than those from fixed-effects models. In this case, six units would be identified as having statistically significant high values for the SMR and five with low values.

This suggests that the estimated intervals are too narrow and, should this method be used, a more appropriate method for estimating them needs to be developed.

Table 5.27 Estimated SMR from random-effects model

Unit	SMR	(95% CI)
1	1.11	(1.00 to 1.35)
2	1.02	(0.92 to 1.25)
3	1.65	(1.50 to 1.95)
4	1.04	(0.90 to 1.33)
5	1.24	(1.12 to 1.52)
6	1.45	(1.30 to 1.80)
7	1.38	(1.25 to 1.69)
8	0.77	(0.70 to 0.93)
9	1.01	(0.88 to 1.30)
10	0.52	(0.46 to 0.65)
11	1.25	(1.13 to 1.55)
12	0.39	(0.36 to 0.49)
13	0.47	(0.41 to 0.59)
14	0.36	(0.32 to 0.44)
15	0.95	(0.86 to 1.17)
16	1.13	(1.01 to 1.44)

5.11 Chapter Summary

In this Chapter various summary measures have been described that can be estimated using logistic regression models. First, the use of the odds ratio was described and the effect of different parameterizations of the units was shown (§5.3). The use of 'deviation from the mean' contrasts was justified as it reduced the influence of large units by using the mean of the logarithm of the odds of the reference units for the comparison.

To overcome potential difficulties in the interpretation of odds ratios, standardized outcomes were introduced in §5.4. Direct and indirect standardization were discussed and appropriate summary statistics described: including the SMR and CMF (§5.5). The use of indirect standardization was proposed for these data (specifically the SMR) because of the problems of direct standardization with small data sets. Section 5.6 comprised an investigation into

proposed methods to estimate a confidence interval for the SMR, and introduced and developed a Bayesian alternative. Whilst an approach employing bootstrap methods demonstrated good coverage, but was shown to be unfeasible for small data sets or where the probability of death was low using the deviation parameterisation preferred in this thesis. An alternative method, proposed by Hosmer & Lemeshow was then advocated for the TNS data.

Mortality difference was briefly discussed as an alternative summary statistic (§5.9). The use of random-effects models was introduced and illustrated (§5.10), but because of the likely decreased sensitivity of such models, compared to fixed effects models, the objectives of this thesis mean that this approach will not pursued further.

Next, the methods proposed in this Chapter were used to investigate the in-unit mortality of the units using more detailed case-mix adjustment. This is described in the next Chapter.

Chapter 6: RISK-ADJUSTED MORTALITY

6.1 Chapter Overview

In this Chapter the methods of Chapter 4 and 5 are brought together to estimate risk-adjusted in-unit mortality rates for infants born at less than 33 weeks gestational age, from 2000 to 2002, and admitted to one of the 16 Trent neonatal units.

Three summary measures of the fit of a logistic regression model are described in §6.3 to enable an assessment of a candidate variable to 'explain' in-unit mortality. The area under the Receiver Operating Characteristic (ROC) curve (§6.3.1) is a measure of the discriminatory ability of a model and the Hosmer & Lemeshow goodness-of-fit χ^2 statistic (§6.3.2) measures its calibration. Cox's measures of calibration and refinement allow an assessment of the calibration of a model on new data (§6.3.3). A more detailed investigation of model fit is left until §6.6.

Section 6.4 comprises an investigation into the association between selected variables recorded by TNS and in-unit mortality. A final model containing all of these potentially important variables is described in §6.5. A reduced model was also estimated, using a stepwise variable selection procedure to identify statistically significant variables (§6.6). The predicted probabilities from the two models are compared in §6.7. The application of the model to more recent TNS data is also discussed (§6.8) and its sensitivity to modelling assumption investigated (§6.9). A Bayesian approach to modelling posterior predictive probabilities is illustrated in §6.10. The problems encountered when a unit has no observed events are discussed (§6.11) and the main points from the Chapter are summarized in §6.12.

6.2 Chapter Introduction

It was noted in Chapter 4 that, for these data, the Trent Neonatal Survey (TNS) did not collect information to allow risk-adjustment using any of the current neonatal mortality risk scores (it has been assumed that CRIB has been superseded by CRIB II). Therefore, here risk-adjustment is achieved through the introduction in a logistic regression model of those variables recorded by TNS and thought to be associated with neonatal mortality. The

omission of some variables identified in published scores, or previous studies, but not collected by TNS, means that risk-adjustment may be incomplete. However, such a situation is not uncommon in medical studies, as it is difficult, perhaps impossible, to quantify completely the morbidity of a human. Nevertheless, it may be possible to identify the factors that have a strong association with mortality and many of these variables are likely to be associated with each other, meaning that the relationship between mortality and a missing variable may, in part, be captured by the inclusion of correlated variables.

The variables to be investigated in this Chapter are those thought to be uninfluenced by any care given by the neonatal team. These variables fall into three broad categories; characteristics of the infant, perinatal factors and antenatal factors.

Gestational age at birth is recognised as a very strong predictor of mortality (§6.4.1), with decreasing mortality with increasing gestational age. Since gestational age is also likely to be associated with other variables, there is the potential for the relationships between mortality and other variable to be confounded by gestational age or for gestational age to be an effect modifier. Therefore, each variable will first be investigated on its own and then with gestational age at birth.

The relationship between the variable of interest and in-unit mortality will be quantified by estimated odds ratios, with Wald 95% confidence intervals (§3.4). The heterogeneity of the effect across NICUs will be tested by adding an indicator variable representing NICU of care and an interaction between NICU and the variable of interest into the logistic model. This will be reported using the p-value for the interaction term. Finally, the estimated standardized mortality ratios for each unit will be reported, obtained using the deviation parameterized model (§5.5.2), and confidence intervals estimated using the method proposed by Hosmer and Lemeshow (1995) described in §5.8.6.

6.3 Summary Measures of Model Fit

With any model, it is important to investigate how well the model fits the data used. Such investigations can be put into one of two broad categories: summary measures of goodness-of-fit and diagnostic approaches looking at the influence of individual observations. Suggested summary measures of model checking are discussed here, whereas model diagnostics are considered in §6.6.

There are three main summary measures used to investigate the fit of a logistic regression model. The Receiver Operating Characteristic (ROC) Curve is used to quantify the discriminatory ability of the model, that is, its ability to distinguish between deaths and survivors. The Hosmer & Lemeshow goodness-of-fit test is used to investigate the calibration of the model. Cox's measures of calibration and refinement allow a more detailed assessment of the calibration of a model but only on new data. Each of these is described in more detail.

6.3.1 Receiver Operating Characteristic (ROC) curve

A plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at all cut-off values is a useful, and well-used, guide to the discriminatory ability of a model. Such a curve is known as a Receiver Operating Characteristic Curve, or, more usually, ROC curve.

The area under such a curve (A_{ROC}) is equivalent to the proportion of times that, given all possible pairings of one survivor and one death, the model would predict a higher probability of death for the observed death than for the observed survivor.

As a guide to interpreting the values of the area under a ROC curve, the following categories have been suggested (Hosmer & Lemeshow, 2000:162):

$A_{ROC} = 0.5$	this suggests no discrimination
$0.7 \leq A_{ROC} < 0.8$	acceptable discrimination
$0.8 \leq A_{ROC} < 0.9$	excellent discrimination
$A_{ROC} \geq 0.9$	outstanding discrimination

Although not explicitly stated by the authors, this grading implies that a value of A_{ROC} of less than 0.7 should be considered as unacceptable discrimination.

6.3.2 Hosmer & Lemeshow goodness-of-fit χ^2 statistic

The calibration of a model is a measure of its ability to predict the observed outcome rates. The approach proposed by Hosmer & Lemeshow (1980) to quantify the calibration of a logistic regression model involves dividing the observations into g groups according to the value of their predicted probabilities. The observed and expected number of deaths can then be compared across all groups. The proposed statistic is given by (Hosmer & Lemeshow, 2000:148):

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(\sum_{i=1}^{n_k} (d_i - \hat{\pi}_i) \right)^2}{\sum_{i=1}^{n_k} \hat{\pi}_i (1 - \hat{\pi}_i)} \qquad (6.1)$$

The number of groups, g, is usually 10 but can take a smaller value when there are less than ten variable patterns. It has been shown that, when the fitted model is correct, $\hat{C}$ is approximated by the χ_{g-2} distribution (Hosmer and Lemeshow, 1980). Hence, large values of $\hat{C}$, and small p-values, indicate a lack of fit of the model.

There is evidence that the value of the test statistic depends on the cut-offs used and may not detect certain types of lack-of-fit (Hosmer *et al,* 1997). However, the test will identify major problems, is simple, easy to explain and available in logistic regression routines in most statistical packages, for example the SAS `PROC LOGISTIC` option `LACKFIT` (SAS Institute Inc., 1999:1961-1962).

6.3.3 Cox's measures of calibration and refinement

A model-based method to assessing the calibration of a predictive model was first proposed by Cox (1958) and can be informative in quantifying the calibration of a model in data other than those used to derive the model parameters. In this approach, a logistic regression model is estimated by regressing the logit of the observed mortality on the logit of the expected probability of death:

$$\log_e \left(\frac{d_i}{1 - d_i} \right) = \alpha + \beta . \log_e \left(\frac{\pi_i}{1 - \pi_i} \right) \qquad (6.2)$$

Hence, α is a measure of the overall calibration of the model when $\beta = 1$, and a measure of the calibration at $\pi = 0.5$ otherwise.

The parameter β is known as the **refinement parameter**. If $\beta > 1$ then the π_i show the right direction but do not vary enough; if $0 < \beta < 1$ the π_i vary too much; if $\beta < 0$ the π_i show the wrong direction (Miller *et al*, 1991).

As well as the numerical estimates for α and β, three hypothesis tests have been put forward to assess the fit of the model (Miller *et al*, 1993):

(i) H_0: $\alpha = 0$, $\beta = 1$, an overall test of the predictions;

(ii) H_0: $\alpha = 0 | \beta = 1$, a test of calibration given correct refinement;

(iii) H_0: $\beta = 1|\alpha$, a test of refinement given correct calibration.

These hypotheses can be tested using likelihood ratio tests and for this thesis the SAS macro proposed by Miller *et al* (1993) was adapted and used.

This methodology is only applicable when the model is being applied to data other than that used to derive the model. The model is, by definition, calibrated to the data used for the modelling process and, therefore, $\alpha = 0$ and $\beta = 1$ in this case.

6.4 Potential Candidate Variables

The Trent Neonatal Survey collects information on many variables (Appendix A). Those variables thought to be uninfluenced by neonatal care were identified and a sub-group of these were to have been selcted for further investigation where there was evidence of an association with neonatal mortality (either from previous studies or from clinical knowledge). However, in practice no variables were excluded at this stage as all identified variables were felt to be potentially associated with mortality.

The candidate variables can be divided into three broad groups. The first of these are the characteristics of the infant: gestational age at birth, birth weight, Apgar score, ethnic origin, base excess and sex. The second group comprises perinatal factors: mode of delivery, antenatal steroid use, infection and fetal distress. The final group of variables are those associated with antenatal factors: mother's age, socio-economic status and mother's gravidity. Variables that can be influenced by treatment or clinical decision by the neonatal team were not considered; for example FiO_2, length of ventilation.

Further details of these investigations are given in Appendix G.

6.4.1 Gestational age at birth

There is known to be a strong monotonic relationship between gestational age at birth of an preterm infant and neonatal mortality (Verloove-Vanhorick *et al,* 1986). For TNS the following hierarchy was used to estimate gestational age (Bohin *et al,* 1999):

i) Mother certain of her dates (most reliable);

ii) Early dating scan

iii) Late dating scan;

iv) Postnatal examination (least reliable).

The linear relationship between gestational age and in-unit mortality was given by:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_G . gest_i$$

$$\hat{\beta}_0 = 16.16 \quad \text{(s.e. 0.88)}$$

$$\hat{\beta}_G = -0.66 \quad \text{(s.e. 0.03)}$$

where: *gest* = gestational age at birth in completed weeks.

($A_{ROC} = 0.881$: $\hat{C} = 3.70 \sim \chi^2_5$, p = 0.59).

There was no statistical evidence that the relationship between gestational age at birth and death before discharge for NICU differed by unit: $\chi^2_{df=15}$=10.19; p = 0.81.

6.4.2 Sex

There has been shown to be a difference in short-term mortality between newborn boys and girls, with boys showing a higher risk of death (Office of National Statistics, 2003b; Effer *et al,* 2002; Larroque *et al,* 2004; Stevenson *et al,* 2000; Shankaran *et al,* 2002; Italian Collaborative Group on Preterm Delivery, 1988).

In the TNS data the in-unit mortality rates were very similar between the sexes:

odds ratio (male vs. female) = 1.01 (95% CI: 0.79 to 1.30); p = 0.91.

After adjusting for gestational age at birth, there was still no evidence that the value of the odds ratio differed from unity:

odds ratio (male vs. female) = 1.18 (95% CI: 0.88 to 1.58); p = 0.27

($A_{ROC} = 0.882$: $\hat{C} = 5.43 \sim \chi^2_7$, p = 0.61)

There was also no evidence for a gestational age-by-sex interaction: p = 0.21, and no evidence that the odds ratios varied across the neonatal units; p = 0.93.

6.4.3 Birth weight

The weight of an infant at birth is known to be associated with its probability of survival (Alberman, 1991). However, it has long been recognised that birth weight in itself is inadequate for predicting mortality (Van Den Berg and Yerushalmy, 1966). Rather, it is the

rate of growth in conjunction with gestational age, i.e. birth weight for gestational age, that is more informative (Coory, 1997).

To investigate any association between birth weight for gestational age and mortality a model including birth weight, gestational age and sex was constructed. The use of fractional polynomials allowed a wide range of possible functions. Using such an approach none of the interactions were statistically significant at the 10% significance level and the final model was:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_S.sex_i + \hat{\beta}_G.gest_i^{-2} + \hat{\beta}_W.weight_i^{-1}$$

$\hat{\beta}_0$ = -9.79 (s.e. 0.48)

$\hat{\beta}_S$ = 0.36 (s.e. 0.16)

$\hat{\beta}_G$ = 3341.91 (s.e. 439.53)

$\hat{\beta}_W$ = 2763.44 (s.e. 299.56)

(A_{ROC} = 0.897: $\hat{C}$ = 6.10 ~ χ^2_8, p = 0.64)

There was no evidence that the relationship between birth weight and in-unit mortality differed between the units: p = 0.60 for an interaction between the inverse of birth weight and unit of care.

The fractional polynomial model was used in this Section, but when more complex models are investigated later in this Chapter, gestational age and birth weight will be included using the *raw data* approach to allow the easier introduction of interactions with other variables.

6.4.4 APGAR score

The Apgar score was originally derived as a simple neonatal morbidity scoring system (Apgar, 1953) and was described in §4.4.8. There is evidence for an association between low Apgar score and increased mortality, including in preterm infants (Casey *et al,* 2001; Weinberger *et al,* 2000). Apgar scores are usually derived at two time points: one minute after birth and again at five minutes. However, Apgar score at five minutes is unsuitable to be included in a model to investigate the quality of care as it is likely to be influenced by early neonatal care.

There was evidence of an interaction between Apgar score at one minute and gestational age at birth:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_A.apgar1_i + \hat{\beta}_G.gest_i + \hat{\beta}_{GA}.gest_i.apgar1_i$$

$\hat{\beta}_0$ = 11.29 (s.e. 2.06)

$\hat{\beta}_A$ = 0.70 (s.e. 0.39)

$\hat{\beta}_G$ = -0.44 (s.e. 0.08)

$\hat{\beta}_{GA}$ = -0.034 (s.e. 0.015)

where: *apgar1* = Apgar Score at 1 minute

gest = gestational age at birth in completed weeks.

(A_{ROC} = 0.900: $\hat{C}$ = 15.99 ~ χ^2_8, p = 0.043)

The value of the *C*-statistic suggests that there is some evidence that this model is a poor fit to the data, but there was no statistical evidence for a non-linear relationship. There was also no evidence of different relationships between Apgar score at one minute and outcome between the units (p = 0.90).

When the model with Apgar score at one minute and gestational age was used to indirectly standardize the in-unit mortality, there was a problem with Unit 9. The only observed death at this unit had a missing Apgar score and was excluded from the model. As there were then no observed deaths for Unit 9, there was quasi-complete separation of the data (§3.4) and the estimates became unstable and had large estimated standard errors. To solve this problem Unit 9 was excluded from the analysis, although this problem is further investigated in §6.9.2.

6.4.5 Ethnic origin

The relationship between ethnicity and neonatal mortality is unclear: e.g. Iyasu *et al* (2002); Cooper *et al* (1993); Singh *et al* (1997); Berman *et al* (2001); Singh *et al* (1997); Berman *et al* (2001); Iyasu *et al* (2002).

In this thesis, the Asian group has been relabelled as 'South Asian' to emphasise the fact that those categorised as Asian are from families originating in South Asia, more particularly from the Indian sub-continent. Other Asian groups, such as Chinese or Filipino, are categorised by TNS as 'Other'.

There was evidence for a difference in mortality between infants of 'European' ethnic origin and those of 'Asian' ethnic origin: overall p = 0.017. However, after the inclusion of gestational age in the logistic regression model, there was no longer evidence for a difference

between the ethnic groups: p = 0.19. There was also no evidence for an interaction with gestational age (p = 0.85), nor for differences in the relationship between the neonatal units (p = 0.99).

6.4.6 Congenital anomalies

Although infants with lethal congenital anomalies have been excluded from all of these analyses, infants with anomalies not thought to be inevitably lethal have been included. High rates of admissions of infants with congenital anomalies are likely to increase in-unit mortality (Sankaran *et al,* 2002).

The unadjusted odds ratio of morality was 1.08 (95% CI: 0.64 to 1.82); p = 0.77. With the introduction of gestational age into the model, there was evidence for a quadratic relationship between gestational age and mortality and for interactions between the presence of a congenital malformation and both linear and quadratic terms for gestational age:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_C.conmal_i + \hat{\beta}_G.gest_i + \hat{\beta}_{CG}.conmal_i.gest_i + \hat{\beta}_{GG}.gest_i^{\,2} + \hat{\beta}_{CGG}.conmal_i.gest^2$$

$\hat{\beta}_0 = 58.62$ (s.e. 41.41)

$\hat{\beta}_C = 74.39$ (s.e. 37.65)

$\hat{\beta}_G = 5.15$ (s.e. 2.98)

$\hat{\beta}_{CG} = -5.74$ (s.e. 2.69)

$\hat{\beta}_{GG} = -0.11$ (s.e. 0.05)

$\hat{\beta}_{CGG} = 0.11$ (s.e. 0.05)

where: $conmal = \begin{cases} 1 & \text{if malformation present} \\ 0 & \text{if no malformation present} \end{cases}$

gest = gestational age at birth in completed weeks.

($A_{ROC} = 0.891$: $\hat{C} = 7.15 \sim \chi^2_5$, p = 0.21)

The addition of an interaction term into the logistic model showed no statistical evidence that the odds ratio varied across the neonatal units (p = 0.99).

6.4.7 Base excess

For TNS, the maximum base excess in the first 12 hours of life is recorded. Abnormal base excess has been shown to be associated with neonatal mortality in preterm infants (The International Neonatal Network, 1993; Maier *et al,* 1997; Garcia *et al,* 2000; Parry *et al,* 2003b).

Seven hundred and forty eight infants (24.7%) had missing values for base excess, of whom 17 (2.3%) died. While the true reason measurements are missing from TNS is unknown, anecdotal evidence suggests that in most cases base excess was not measured when it was felt likely to be in the normal range (Field, D.J.: Personal communication). In this case, it may be appropriate to substitute the missing values with a 'normal' value. Such an approach is commonly used with published risk-adjustment scores, for example PIM (Shann *et al,* 1997), SNAP (Richardson *et al,* 1993), MMPS (Daley *et al,* 1988). Using this assumption, it was possible to categorise all of the observations according to their estimated base excess by putting those with missing values into the 'normal group'. The groups used here were those from the original CRIB score: >-7.0, -7.0 to -9.9, -10.0 to –14.9 and ≤ 15.0 mmol/L. There was evidence of a difference in mortality rates between the groups ($p < 0.0001$), with increasing mortality with decreasing maximum recorded base excess. The inclusion of gestational age showed evidence for an interaction between gestational age and maximum base excess group: $p = 0.0003$:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_{B2}.baseexcess2_i + \hat{\beta}_{B3}.baseexcess3_i + \hat{\beta}_{B4}.baseexcess4_i + \hat{\beta}_G.gest_i + \hat{\beta}_{B2G}.baseexcess2_i.gest_i + \hat{\beta}_{B3G}.baseexcess3_i.gest_i + \hat{\beta}_{B4G}.baseexcess4_i.gest_i$$

$\hat{\beta}_0 = 18.19$ (s.e. 1.50)

$\hat{\beta}_{B2} = -2.81$ (s.e. 2.63)

$\hat{\beta}_{B3} = -7.16$ (s.e. 2.37)

$\hat{\beta}_{B4} = -7.06$ (s.e. 2.69)

$\hat{\beta}_G = -0.76$ (s.e. 0.06)

$\hat{\beta}_{B2G} = 0.12$ (s.e. 0.10)

$\hat{\beta}_{B3G} = 0.32$ (s.e. 0.09)

$\hat{\beta}_{B4G} = 0.135$ (s.e. 0.10)

where: $baseexcess.2 = \begin{cases} 1 & \text{if maximum base excess} = -7.0 \text{ to } -9.9 \\ 0 & \text{else} \end{cases}$

$baseexcess.3 = \begin{cases} 1 & \text{if maximum base excess} = -10.0 \text{ to } -14.9 \\ 0 & \text{else} \end{cases}$

$baseexcess.4 = \begin{cases} 1 & \text{if maximum base excess} \leq 15.0 \\ 0 & \text{else} \end{cases}$

gest = gestational age at birth in completed weeks.

($A_{ROC} = 0.911$: $\hat{C} = 1.10 \sim \chi^2_8$, p = 0.98)

The introduction of interaction terms between the units and base excess groups showed no improvement in the fit of the model: p = 0.98.

6.4.8 Multiplicity of pregnancy

There has been evidence presented that multiple birth is a risk factor amongst extremely low birth weight infants (501-1000g) (Shankaran *et al,* 2002). However, there is also evidence that twins have better gestational age specific neonatal survival rates than singletons (Kiely, 1998).

Since there were a relatively small number of triplets, and only three deaths, a dichotomous variable was used: singleton or multiple birth.

The TNS data showed decreasing mortality for multiple births, although this difference was not statistically significant: p-value = 0.17. The effect of gestational age at birth was investigated using a logistic regression model and there was evidence of a gestational age-by-multiplicity interaction (p = 0.0065):

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_M .multiple_i + \hat{\beta}_G .gest_i + \hat{\beta}_{MG} .multiple_i .gest_i$$

$\hat{\beta}_0 = 14.95$ (s.e. 0.95)

$\hat{\beta}_M = 7.19$ (s.e. 2.68)

$\hat{\beta}_G = -0.62$ (s.e. 0.04)

$\hat{\beta}_{MG} = -0.28$ (s.e. 0.10)

where: $multiple = \begin{cases} 1 & \text{if multiple birth} \\ 0 & \text{if singleton birth} \end{cases}$

gest = gestational age at birth in completed weeks

(A_{ROC} = 0.885: $\hat{C}$ = 4.97 ~ χ^2_8, p = 0.66)

There was evidence that infants from multiple pregnancies do worse that singletons if born before about 26 weeks gestational age, but appear to do better if born after this time. There was no evidence that the relationship between mortality and multiple birth was different between the units, after adjusting for gestational age (p = 0.99).

6.4.9 Socio-economic status

The postcode of the mother's place of residence was recorded by TNS and could be used to investigate any association between area-based socio-economic deprivation and in-unit neonatal morality.

The area-based deprivation scoring systems chosen was the Index of Multiple Deprivation 2000 (IMD) published by the Department of the Environment, Transport and the Regions.

The odds ratio for mortality for a unit increase in IMD was 1.00 (95% CI: 0.99 to 1.01), p = 0.49. When gestational age was included in the model (no evidence for an interaction by gestational age: p = 0.44) the odds ratio for a unit increase in IMD was 1.00 (95% CI: 0.99 to 1.01), p = 0.89 (A_{ROC} = 0.886: $\hat{C}$ = 5.55 ~ χ^2_7, p = 0.59). There was no evidence that this relationship differed between units: p = 0.44.

6.4.10 Antenatal corticosteroids

The use of antenatal corticosteroids prior to preterm birth has long been known to reduce subsequent respiratory distress syndrome (RDS) (Liggins and Howie, 1972) and, therefore, neonatal mortality (Crowley, 2003).

The estimated odds ratio for mortality, for infants of mothers given antenatal corticosteroids compared to those who were not, was 0.62 (95% CI: 0.47 to 0.81), p = 0.0004. When gestational age was included in the model (no evidence for an interaction by gestational age: p = 0.46) the estimated odds ratio was 0.63 (95% CI: 0.45 to 0.87), p = 0.0047 (A_{ROC} = 0.883: $\hat{C}$ = 6.01 ~ χ^2_6, p = 0.42). There was no evidence that this relationship differed between units: p = 0.63.

6.4.11 Intrapartum monitoring

The term **fetal distress** is an often-used description of problems during birth, although it has no clear definition. In general, it describes the situation where the fetus is deprived of oxygen during labour or delivery, although this is often called acute fetal distress to distinguish it from sustained hypoxia during the pregnancy, which is usually termed chronic fetal distress.

A number of different approaches have been advocated to try to detect fetal distress (Mead, 1996) and six variables are collected in TNS to try to identify those deliveries where fetal distress occurred: 'fetal distress', 'CTG abnormality', 'Doppler abnormality', 'abnormal scalp pH', meconium present' and 'other'. It has been suggested that such indirect measures that try to indicate hypoxia are poor predictors (Low *et al,* 1995b). Therefore, an infant with any of the indicators recorded as abnormal was assumed to have experienced fetal distress.

The estimated odds ratio for mortality was 0.90 (95% CI: 0.70 to 1.16), p = 0.42. When gestational age at birth was included in the model there was evidence for a fetal distress by gestational age interaction (p = 0.016).

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_D.distress_i + \hat{\beta}_G.gest_i + \hat{\beta}_{DG}.distress_i.gest_i$$

$\hat{\beta}_0$ = 18.44 (s.e. 1.25)

$\hat{\beta}_D$ = -3.98 (s.e. 1.91)

$\hat{\beta}_G$ = -0.76 (s.e. 0.05)

$\hat{\beta}_{DG}$ = 0.17 (s.e. 0.07)

(A_{ROC} = 0.886: $\hat{C}$ = 4.80 $\sim \chi^2_7$, p = 0.68)

where: $distress = \begin{cases} 1 & \text{if sign of fetal distress recorded} \\ 0 & \text{otherwise} \end{cases}$

gest = gestational age at birth in completed weeks

There was no evidence that the relationship between reported fetal distress and mortality differed amongst the units (p = 0.96). The estimated functions suggest that fetal distress was associated with mortality for infants born at 25 to 30 weeks gestational age (Figure G.33).

6.4.12 Mode of delivery

There is conflicting evidence on whether caesarean section reduces mortality for preterm infants (Penn and Ghaem-Maghami, 2001). Guidelines from the National Institute for Clinical Excellence (NICE) state that, because of the uncertainty over the outcomes from planned caesarean section for preterm births, planned caesarean sections *"should not be routinely offered outside a research context"* (National Institute for Clinical Excellence, 2004:1.2.3.1).

In order to obtain reasonably sized groups, the methods of delivery recorded by TNS were combined into three clinically homogeneous groups: vaginal delivery, labouring caesarean section and non-labouring caesarean section (Field, D.J.: Personal communication).

Infants delivered by caesarean section had statistically significant higher rates of survival than those delivered vaginally: overall p-value = 0.0043. However, once gestational age was included in the model (p-value for interaction = 0.25) the situation was reversed: infants born by vaginal delivery had the lowest gestational age specific mortality rates. There was a difference between those deliveries where labour occurred (vaginal and labouring caesarean section) and those where it did not (both non-labouring caesarean section):

Mode of delivery	**Odds ratio**	**(95% CI)**	**p-value**
Vaginal	reference		
CS: labouring	0.59	(0.40 to 0.87)	0.0076
CS: non-labouring	0.70	(0.53 to 0.92)	0.0092

There was no evidence for an interaction between mode of delivery and NICU: p = 0.98.

6.4.13 Mother's age

There is a large body of evidence that the risk of complications during pregnancy, and at delivery, increases with increasing maternal age (Fretts *et al,* 1995; Jolly *et al,* 2000; Temmerman *et al,* 2004). However, the link between neonatal outcomes and maternal age is less clear, with some evidence that there is no increased risk of poor neonatal outcomes with increased maternal age (Berkowitz *et al,* 1990).

There was no evidence for a relationship between a mother's age and the infant's risk of death: estimated odds ratio = 1.00; 95% CI 0.98 to 1.02; p = 0.98. This did not change after

including gestational age at birth in the model: estimated odds ratio = 0.99; 95% CI 0.97 to 1.02; p = 0.54 (A_{ROC} = 0.884: $\hat{C}$ = 10.48 ~ χ^2_8, p = 0.23). There was also no evidence that the relationship between the mothers' ages and infant mortality varied in the different neonatal units: p = 0.25.

6.4.14 Previous obstetric history

The relationship between a mother's obstetric history and the probability of mortality for a subsequent infant is unclear (Bai *et al,* 2002; Billewicz, 1973; Roman *et al,* 1978; Bakketeig and Hoffman, 1979). In this thesis only the previous number of pregnancies (**gravidity**) was considered. To avoid small numbers of observations in each group, the observations were divided into three categories: primagravida, secundigravida, multigravida.

There was statistical evidence for difference in the mortality rates between the groups: p = 0.029. However, once gestational age is included in the model (p-value for interaction = 0.28), there was no evidence for a difference between the groups in gestational age specific mortality rates: p = 0.58.

6.4.15 Maternal or fetal infection

Maternal or fetal infection is known to increase the risk of preterm birth and to increase mortality among preterm infants (Fung *et al,* 2003; Garite and Freeman, 1982; Ernest, 1998).

The observed odds ratio for mortality was 1.23 (95% CI: 0.91 to 1.66), p = 0.18. After adjustment for gestational age (p = 0.72 for interaction between infection and gestational age) the adjusted odds ratio was 0.84 (95% CI: 0.59 to 1.21), p = 0.34 (A_{ROC} = 0.882: $\hat{C}$ = 4.11 ~ χ^2_7, p = 0.77). There was no evidence that the relationship between infection and gestational age differed by unit: p = 0.95.

6.4.16 Summary of risk adjustment using TNS variables

All of the variables discussed above were used in a final model (§6.5). Although some variables did not reach the specified level of statistical significance, it is the clinical, not statistical, significance of the variables that is of interest. For example, there was no statistical evidence for an association between infection and mortality, but adjustment for infection produced 'significant' changes to the estimated SMRs. The influence of selected variables on the estimated SMRs was shown individually, and after adjustment for gestational

age, but their joint influence was of more interest. A reduced model will be estimated in order to try to identify a more parsimonious model (§6.6).

Other potentially important factors have been suggested in the medical literature but are not recorded by TNS and, therefore, not included in any modelling here. These variables include sibship size (Bakketeig and Hoffman, 1979) and mean blood pressure (Jain and Fleming, 2004). The quality of antenatal care has also been proposed as a predictor of neonatal mortality (Vintzileos *et al,* 2002). It was advocated that the absence of antenatal care was associated with higher neonatal death, after adjusting for maternal age, birth weight and gestational age. However, this relationship was greater among term infants, ≥36 weeks: relative risk for death = 2.1 (95% CI: 1.8 to 2.4), than among preterm infants, 24-35 weeks: relative risk = 1.2 (95% CI: 1.1 to 1.3). It is unknown whether these variables have an effect over and above the variables in the model.

6.5 Full Model

It has been argued that a predictive model should include all variables thought to be clinically important (Healthcare Quality and Analysis Division, 2002; Center for Health Services Research in Primary Care, 1996) For this reason the first model contains all of the variables investigated in the previous Section.

6.5.1 Model parameters

Plausible clinical interactions between variables discussed in the previous Section were included where they were statistically significant at the 10% level. Birth weight, gestational age and sex were included using the first model discussed in Appendix G.3, i.e.:

$$g_D = \beta_0 + \beta_S.sex + \beta_G.gestation + \beta_W.birthweight + \beta_{WW}.birthweight^2 .$$

This allowed gestational age to be included in the model as a linear term, enabling interaction with other variables to be included in a simple manner. There were 282 observations with missing data that were omitted from the model, leaving 2502 survivors and 241 deaths. The parameter estimates are shown in Table 6.1.

Table 6.1 Parameter estimates from full model

Variable	Group	$\hat{\beta}$	s.e.	P-value
Intercept		7.13	14.42	
Gestational age (week)	Linear	0.24	1.11	
	Quadratic	-0.013	0.021	
Sex	Female	Reference		0.013
	Male	0.45	0.18	
Birth weight (g)		-0.0085	0.0013	
		0.0000025	0.0000005	< 0.0001
Apgar at 1 minute		0.021	0.46	
Apgar*gestational age		-0.0053	0.0173	0.76
Ethnicity	European	Reference		0.78
	South Asian	0.13	0.31	
	Other/unknown	-0.15	0.30	
Congenital malformation	None	Reference		
	Present	121.3	56.0	
Congenial malformation * gestational age	None	Reference		
	Present	-9.12	4.00	
Congenial malformation * gestation^2	None	Reference		0.016
	Present	0.17	0.07	
Base excess	> -7.0 (mmol/L)	Reference		
	-7.0 to –9.9	-2.00	3.00	
	-10.0 to –14.9	-7.98	2.83	
	≤ -15.0	-8.05	3.23	
Base excess * gestational age	> -7.0 (mmol/L)	Reference		0.0020
	-7.0 to –9.9	0.087	0.11	
	-10.0 to –14.9	0.34	0.11	
	≤ -15.0	0.37	0.12	
Multiple birth	Singleton	Reference		
	Multiple	5.25	3.11	
Multiple birth * gestational age	Singleton	Reference		0.10
	Multiple	-0.19	0.12	
IMD		-0.0062	0.0054	0.25
Corticosteroids	No	Reference		0.073
	Yes	-0.36	0.20	
Fetal distress	No	Reference		
	Yes	-2.80	2.58	
Fetal distress * gestational age	No	Reference		0.27
	Yes	0.11	0.096	
Mode of delivery	Vaginal	Reference		
	CS: labouring	1.73	3.87	
	CS: non-labour	2.37	3.21	
Mode * gestational age	Vaginal	Reference		0.82
	CS: labouring	-0.065	0.15	
	CS: non-labour	-0.071	0.12	
Mother's age (year)		-0.012	0.016	0.43
Gravidity	Prima	Reference		0.20
	Secund	0.22	0.25	
	Terce	0.40	0.22	
Infection	No	Reference		
	Yes	-0.093	2.56	
Infection * gestational age	No	Reference		0.92
	Yes	0.010	0.097	

6.5.2 Model checking

The validity of the model was examined. First, the goodness-of-fit of the model to the current data was investigated using assessment by deletion: i.e. deleting each observation in turn and re-estimating the model parameters (jacknife). Diagnostic plots showing the resultant changes in the values of the Pearson chi-square statistic, deviance and parameter estimates were produced and inspected. A deletion approach was also used to obtain jacknife predicted probabilities of the individual observations. Finally, the calibration and discrimination of the model were investigated.

Assessment by deletion

Hosmer and Lemeshow (2000:176) have suggested various diagnostic plots that are useful for logistic regression models and each of these was investigated for the model.

The first approach uses the change in the value of the **Pearson chi-square statistic** after the deletion of an observation (ΔX^2). The value of the change is plotted it against the predicted probability for each observation. The Pearson chi-square statistic is the sum of the Pearson residuals where the Pearson residual for observation *i* is given by:

$$r_i = \left(\frac{d_i - \hat{\pi}_i}{\hat{\pi}_i \left[1 - \hat{\pi}_i \right]} \right)$$

It can further be shown (Hosmer & Lemeshow, 2000:174) that, for observation *i*, the change in the value of the Pearson chi-square statistic is given by:

$$\Delta X_i^2 = \frac{r_i^2}{(1 - h_i)}$$

where: h_i is the i^{th} diagonal element of the **H** matrix (i.e. the leverage).

The reason for using ΔX_i^2 instead of r_i is that positive values of r_i are from observations where d_i = 1, and negative values are from observations where d_i = 0. Therefore, the sign of the residual does not carry any extra information and squaring the value emphasizes any lack of fit. Such plots show observations with predicted probabilities far from the observed outcome. These are sometimes plotted (as in Figure 6.1) with the size of the plotting symbol proportional to the standardized change in the value of the parameter estimates after the deletion of observation *i* ($\Delta\hat{\boldsymbol{\beta}}_i$). $\Delta\hat{\boldsymbol{\beta}}_i$ is given by:

$$\Delta\hat{\boldsymbol{\beta}}_i = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i^1\right)' \left(\mathbf{X}'\mathbf{V}\mathbf{X}\right)^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i^1\right)$$

where: **V** is the *m* x *m* diagonal matrix: $diag\{\hat{\pi}_i(1-\hat{\pi}_i)\}$;

X is the data matrix

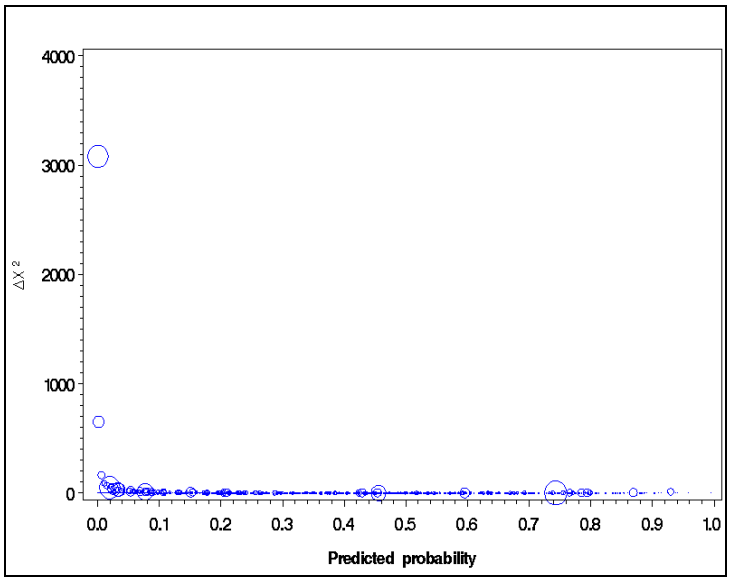
This can be shown (Pregibon, 1981) to be equivalent to:

$$\Delta\hat{\boldsymbol{\beta}}_i = \frac{r_i^2 h_i}{(1-h_i)^2}$$

This plot is useful in indicating the influence of any outstanding observation.

The two observations noted in Figure 6.1 as being outliers were infants who died, but for whom the model suggested good prognoses: predicted probabilities of death of 0.00032 and 0.0015. Both of these infants were born at 31 weeks gestational age, at around 1600g birth weight, had high Apgar scores at one minute of life (7 & 9) and 'good' values for all other variables. These values were check on the relevant TNS forms and found to have been entered onto the database correctly.

Figure 6.1 Change in Pearson chi-square statistic



The second diagnostic plot shows the change in **deviance** after the deletion of an observation, ΔD (Figure 6.2). The deviance residual for observation *i* is given by:

$$d_i = \begin{cases} -\sqrt{2\left|\ln(1-\hat{\pi}_i)\right|} & \text{if } d_i = 0 \\ \sqrt{2\left|\ln(\hat{\pi}_i)\right|} & \text{if } d_i = 1 \end{cases}$$

and the deviance D is given by the sum of the deviance residuals:

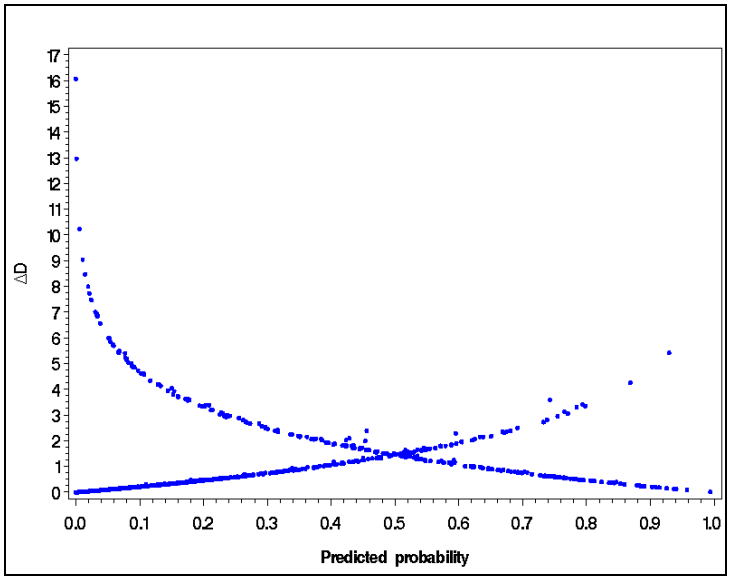
$$D = -2\sum_{i=1}^{N}\left(d_i \ln\left[\frac{\hat{\pi}_i}{d_i}\right] + [1-d_i]\ln\left[\frac{1-\hat{\pi}_i}{1-d_i}\right]\right)$$

The change in deviance can be shown to be approximated by (Hosmer & Lemeshow, 2000:174):

$$\Delta D_i = \frac{d_i^2}{(1-h_i)}$$

The two observations with values of ΔD_i greater than 12 are the same as those noted in Figure 6.1. The third observation with $\Delta D_i > 10$ was an infants born at 30 weeks, with good indicators for survival as measured by the model ($\hat{\pi}$ = 0.0060), who died before discharge.

Figure 6.2 Change in deviance



The third step is to consider the effect of removing each observation in turn on each **parameter estimate**, *DFBETA*. In such situations, rather than refitting the model n times (where n is the number of observations) a one-step approximation to the parameter estimates is generally used (SAS Institute Inc., 1999:1957). The MLE for the model parameters estimated without observation i ($\hat{\boldsymbol{\beta}}_i^1$) is given by:

$$\hat{\boldsymbol{\beta}}_i^1 = \hat{\boldsymbol{\beta}} - \frac{(d_i - \pi_i)}{(1 - h_i)} \hat{\mathbf{V}}_\beta \mathbf{x}_i \qquad (6.3)$$

where: $\hat{\boldsymbol{\beta}}$ is the MLE of parameter vector $(\beta_0, \beta_1, ..., \beta_P)$ using all observations;

$\hat{\mathbf{V}}_{\boldsymbol{\beta}}$ is the estimated covariance matrix of $\boldsymbol{\beta}$.

The change in parameter estimates from removing observation *i* from the model is given by:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i^1 = \boldsymbol{\Delta\beta}_i^1 = \frac{(d_i - \pi_i)}{(1 - h_i)} \hat{\mathbf{V}}_\beta \mathbf{x}_i \qquad (6.4)$$

Hence, if $\Delta_p \beta_i^1$ is the p^{th} component of the one-step difference (*6.3*) then the standardized difference for parameter *p* after deletion of observation *i* is given by:

$$DFBETAp_i = \frac{\boldsymbol{\Delta}_p \beta_i^1}{\hat{\sigma}(\beta_p)}$$

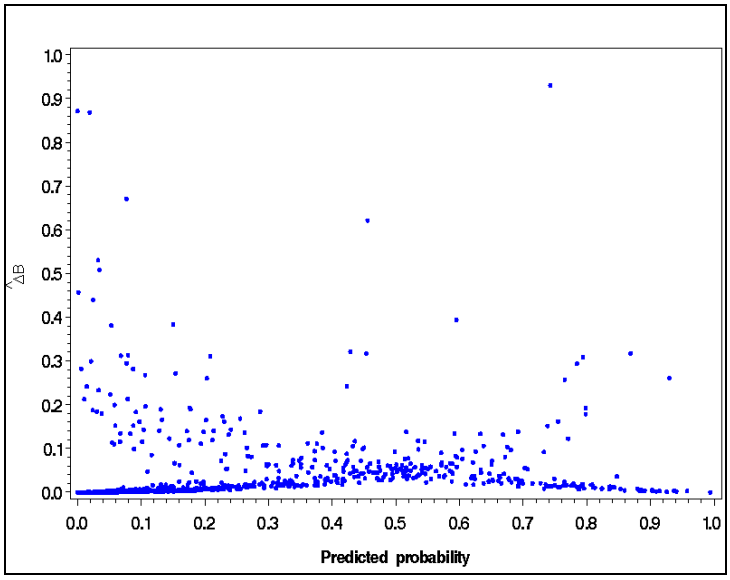
where: $\hat{\sigma}(\beta_p)$ is the estimated standard error of parameter *p* from the full model.

These values were plotted for each parameter in the model (Appendix H). All of the values were less than 1.0, the minimum value suggested as signifying that an observation has a significant effect on the value of a parameter estimate (Hosmer & Lemeshow, 2000:180).

However, here the value of the parameter estimates themselves were not of primary importance, rather it was the predicted probabilities of death that were of interest. A change in the value in one parameter estimate can be offset by the change in value of one or more of the other parameter estimates, resulting in similar predicted values. This is particularly true when there are interactions in the model, as in the model under consideration here. A more interesting approach may be to consider changes in the vector of parameter estimates as a whole, quantified by $\Delta\hat{\boldsymbol{\beta}}_i$. This plot is shown in Figure 6.3.

There were two observations with large values of $\Delta\hat{\boldsymbol{\beta}}_i$ relative to the other observations (Figure 6.3). However, all took values of less than 1.0 and it has been suggested that observations with values of less than 1.0 are not likely to have a 'significant' effect on the estimation of the values of the parameter estimates (Hosmer & Lemeshow, 2000:180). Nevertheless, it was still of interest to identify the observations with high values, using 0.5 as an arbitrary cut-off.

Figure 6.3 Change in model parameter estimate values



There were seven observations where $\Delta\hat{\boldsymbol{\beta}}_i > 0.5$; six of whom died and one survived to discharge. The infant who survived was born at 26 weeks and 660g with a very poor Apgar score at one minute of life and would have been expected to have a poor prognosis ($\hat{\pi}$ = 0.74). However, the infant was discharged home after 136 days on a neonatal unit. Five of the other infants identified were of 30 to 32 weeks gestational age with good prognosis (as measured by the model: $\hat{\pi}$ of 0.00032 to 0.077) who nevertheless died before discharge. One of these was also identified as an outlier in Figure 6.1 and Figure 6.2. The final observation identified was a small infants (24 weeks and 745g) but with other prognostic factors indicating a mixed prognosis (e.g. presence of congenital malformation and fetal distress, Apgar score of 6 at 1 minute, maximum base excess of -9.4): $\hat{\pi}$ = 0.46 . Inspection of the individual variable DFBETAs for this infant (Appendix H) showed relatively high values for congenital malformation (*DFBETA* = 0.58) and its interaction with gestational age (*DFBETA* = -0.56) and the square of gestational age (*DFBETA* = 0.54).

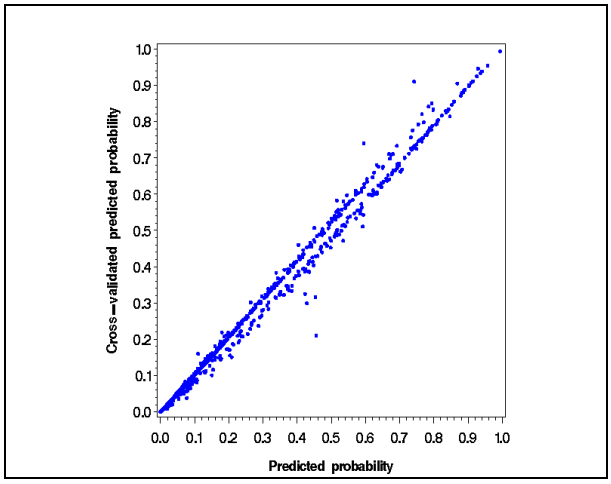
The recorded observations for all of these infants were checked against the TNS forms and no discrepancies were found. As these were genuine observations and it was decided that they should be included in all modelling.

Internal model validation

While the influence of individual observations on the value of parameter estimates is of interest, of more importance in this thesis is the stability of the individual predicted probabilities of death before discharge. This can be investigated using internal cross

validation of the model but this can be undertaken in several ways. Often the approach is to split the data into groups, either at random or by using some other methods, for example odd and even identification numbers (Cole *et al,* 1991). The values of the model parameters are then estimated using part of the data ('training set'). These estimates are then applied to the other part of the data ('test set') and which is then inspected for predictive ability. However, such data splitting methods are weak, as the two sets will usually be similar other than for random variation. A more robust approach is to use a 'deletion' approach. This involves excluding a single observation or group of observations (e.g. single unit) in turn. The model parameters are estimated with each observation (or group of observations) deleted and are then used to obtain a predicted probability for the deleted observations. These probabilities are then compared to the predictions from the model that used the whole data. It has been suggested that the 'leave-one-out' approach is superior (Altman and Royston, 2000) and this was the approach taken here. Such jacknife predicted probabilities were estimated using the one-step estimates outlined above (*6.2*) and are shown in Figure 6.4. There was good agreement between the predicted probabilities from the model and the jacknife estimated probabilities.

Figure 6.4 Cross-validated predicted probabilities



Calibration of the model

As the parameter estimates from this model were used to obtain indirectly standardized mortality rates, it is the calibration of the model, its ability to assign 'correct' mortality probabilities, that is of particular importance. There was no evidence of poor calibration from the Hosmer & Lemeshow goodness-of-fit test (§6.3.2): $\hat{C} = 11.69 \sim \chi^2_8$, p = 0.17. An

alternative approach to investigating the calibration of a model, that may offer a deeper insight into the calibration of the model, is to plot calibration curves. This is similar in approach to the Hosmer and Lemeshow goodness-of-fit test (§6.3.2) in that the observations are divided into strata and the observed and expected mortality are compared within each stratum. However, such calibration curves differ in two respects from the usually applied version of the Hosmer and Lemeshow test. First, rather than being divided into approximately equal sized strata, the observations are divided according to the value of the predicted probabilities. Second, the differences between the observed and expected number of deaths are inspected visually to gain a deeper understanding of any deficiencies in the model. These values are plotted and then inspected and compared with a diagonal line representing perfect predictive ability (Rowan *et al,* 1993b). There is obviously uncertainty around the lines in such calibration plots but it is unclear how this uncertainty can be quantified. One possibility is to repeatedly model bootstrap samples of the data but, as has been discussed elsewhere in this thesis, such an approach is computationally intensive and has not been persued here. Therefore, the plots are shown without any indication of the size of the errors.

Figure 6.5 suggests that this model was very well calibrated as the observed and expected number of deaths are very similar across all values of predicted mortality. To investigate whether this property holds over all gestational ages, calibration plots for two groups split by gestational age are shown in Figure 6.6.

Figure 6.5 Calibration plot: full model

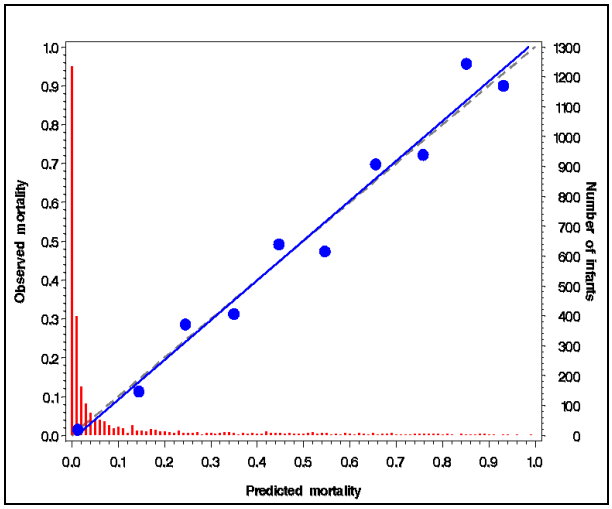
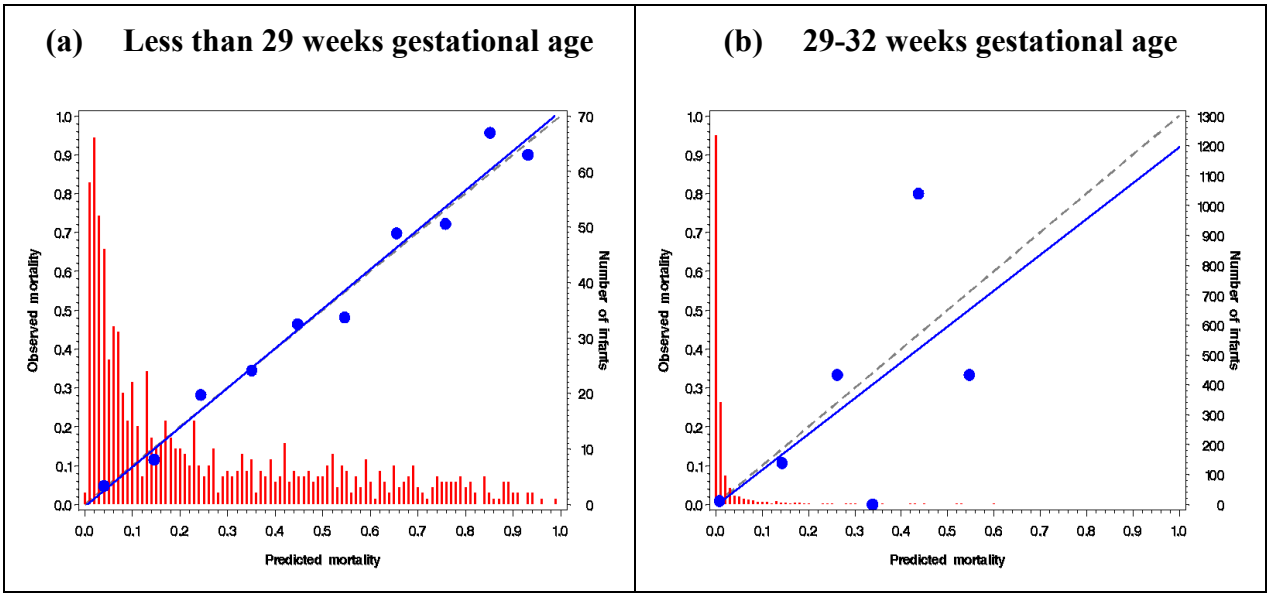
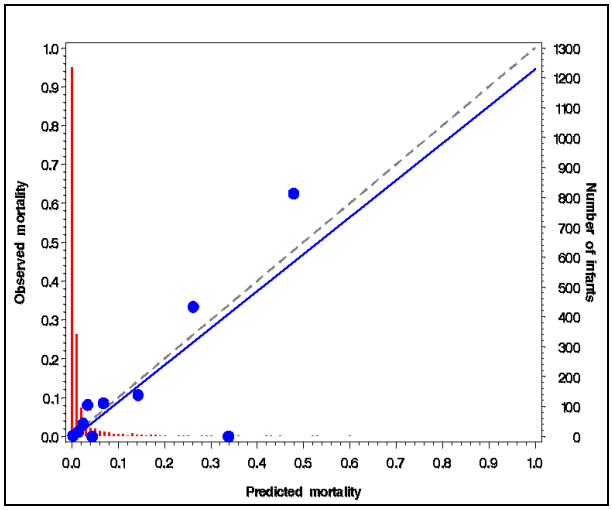


Figure 6.6 Calibration plots by gestational age group



For those infants born at 29 to 32 weeks, although the best-fit line approached the line of no difference (i.e. 45º), some of the points fell far from it (Figure 6.6b). This was likely to be because most of the predicted probabilities are very small for this group (mean = 0.016, maximum = 0.51) and the poor appearance of the calibration plot may, in part, be due to the small number of observations with high predicted probabilities. To reduce the influence of these few observations, the plot was re-drawn using different strata: $(0 \leq \pi < 0.01)$; $(0.01 \leq \pi < 0.02)$; $(0.02 \leq \pi < 0.03)$; $(0.03 \leq \pi < 0.04)$; $(0.04 \leq \pi < 0.05)$; $(0.05 \leq \pi < 0.1)$; $(0.1 \leq \pi < 0.2)$; $(0.2 \leq \pi < 0.3)$; $(0.3 \leq \pi < 0.5)$; $(0.5 \leq \pi \leq 1.0)$ (Figure 6.7).

Figure 6.7 Calibration plot: 29-32 weeks gestational age

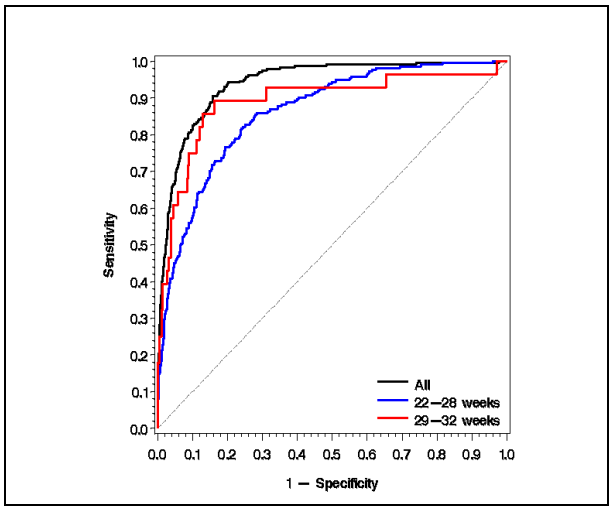


There was still some evidence that the model overestimated mortality. However, the difference between the calibration line and the reference line was small, especially over the range of predicted mortality obtained from the model.

Discrimination of the model

The ability of a predictive model to discriminate between outcomes is usually measured by the area under the Receiver Operator Characteristic (ROC) curve (§6.3.1). ROC curves for the model are shown in Figure 6.8, together with curves for two subsets of infants defined by gestational age.

Figure 6.8 ROC curves for 'full' model



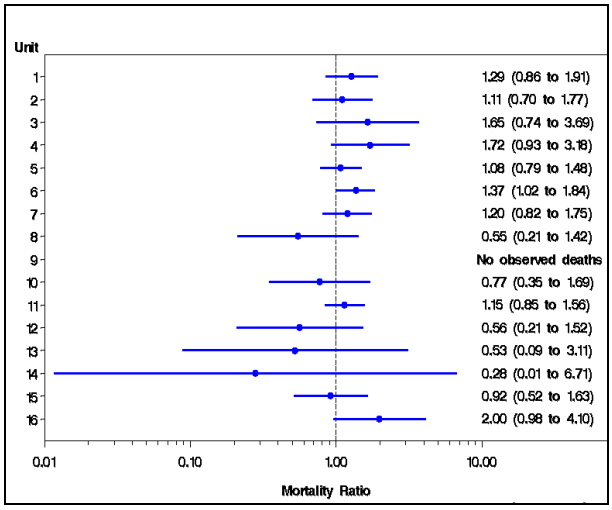
The value for the area under the ROC curve for all of the data was 0.940 (s.e. = 0.007); such a value has been described as 'outstanding' (Hosmer & Lemeshow, 2000:162). When the observations are considered by gestational age the values were 0.861 (s.e. = 0.014) for the 22 to 28 week infants and 0.893 (s.e. = 0.040) for the 29 to 32 week infants, both considered 'excellent'. It was noted that both of these values were less than that for all of the data. This was because gestational age was a very strong predictor of mortality and analysing sub-groups based on gestational age reduced its effect.

All of these values compared well with the values for gestational age alone: all infants $AROC$ = 0.881; 22-28 weeks $AROC$ = 0.752; 29-32 weeks $AROC$ = 0.641.

Estimated Standardized Mortality Ratios

After adjusting for all of the variables, the lower limit of the estimated 95% confidence interval for Unit 6 is still greater than one (Figure 6.9). It is also of interest to note the estimated interval for Unit 16, which now has its lower limit at just below the value one.

Figure 6.9 Estimated Standardized Mortality Ratios: full model



6.6 *Reduced Model*

Some of the variables in the full model, described in the previous Section, did not reach statistical significance at the 10% level. While it is recognised that this does not mean that they are not clinically important, nor that they do not have an impact on the estimated SMRs, the introduction of variables into a logistic regression model increases the variance of the parameter estimates (Robinson and Jewell, 1991). Therefore, it may useful to remove unnecessary variables from the model, to see if the precision of the parameter estimates increases without substantially changing the point estimates.

6.6.1 Model selection method

The process of model selection is important, as different methods can often produce different 'optimal' models (Agresti, 1990:218-219). The first choice of a data driven method to model selection is the 'best subsets' approach, although this can require the estimation of many parameters in many models resulting in problems for some statistical packages (Hosmer and Lemeshow, 2000:134-135). This was the case with these data. It was, therefore, decided to use forward stepwise model selection. The process starts from the null model and terms are added according to the value of the score chi-square statistic. The term with the largest value is included in the model if it is statistically significant. At each step terms may also be removed from the model if their removal, as measured by the Wald test, does not produce a statistically significant change in model fit (SAS Institute Inc., 1999:1945-1946). The process continues until no other term is added to the model or if the term just added is the only one to be excluded. For this analysis the significance levels for inclusion and exclusion were both

set at 10%. To obtain the reduced model, all of the variables in the full model were considered for inclusion, as were quadratic terms for gestational age and birth weight, and all possible two-way and three-way interactions.

Although such data-driven model selection procedures are commonly used (Armitage and Berry, 1994:321-322; Hosmer and Lemeshow, 2000:116), it is acknowledged that their use can induce problems (Harrell, 2001:56-57). One of these arises from the multiple comparisons made during the model selection procedure, leading to p-values that are too small. However, for this analysis it was the accuracy of the predicted values that was of principal importance, rather than the statistical significance of the included variables or the composition of the set of included variables themselves, and so this was not a problem (Bland, 1995:323). While a clinically plausible model may increase the confidence of potential users of the model, this is not an aim of the modelling process here.

6.6.2 Reduced model

All data without missing values for the variables in the model were included. The final 'reduced' model was estimated using 2885 observations (Table 6.2).

Table 6.2 Data in reduced model

Unit	Observed			Missing	
	Infants	Died	(%)	Infants	Died
1	205	21	(10.2)	7	0
2	265	24	(9.1)	18	6
3	38	2	(5.3)	0	0
4	139	5	(3.6)	3	1
5	322	38	(11.8)	11	3
6	372	51	(13.7)	6	3
7	227	22	(9.7)	16	7
8	104	4	(3.9)	20	4
9	32	0	(0)	3	1
10	141	5	93.6)	5	0
11	421	55	(13.1)	24	7
12	190	5	(2.6)	6	0
13	126	2	(1.6)	10	1
14	85	1	(1.2)	5	1
15	122	10	(8.2)	2	0
16	96	6	(6.3)	4	0
Total	**2885**	**251**	**(8.7)**	**140**	**34**

The final model contained terms for gestational age, sex, birth weight (up to quadratic), Apgar score at one minute, base excess and an interaction between base excess and birth weight (Table 6.3). An interaction between maximum base excess and birth weight was the only term included here but not in the full model.

Table 6.3 Parameter estimates: reduced model

Variable	Group	$\hat{\beta}$	s.e.	P-value
Intercept		12.40	1.32	
Gestational age (week)		-0.28	0.06	< 0.0001
Sex	Female	Reference		0.022
	Male	0.40	0.17	
Birth weight (g)	Linear	-0.0096	0.0013	
	Quadratic	0.0000024	0.0000005	< 0.0001
Apgar		-0.13	0.04	0.0004
Base excess	> -7.0 (mmol/L)	Reference		
	-7.0 to –9.9	-0.55	0.76	
	-10.0 to –14.9	-1.32	0.64	
	≤ -15.0	-0.72	0.83	
Base excess * birth weight	> -7.0 (mmol/L)	Reference		0.0002
	-7.0 to –9.9	0.0010	0.0009	
	-10.0 to –14.9	0.0026	0.0007	
	≤ -15.0	0.0029	0.0008	
(A_{ROC} = 0.932: $\hat{C}$ = 6.43 ~ χ^2_8, p = 0.60)				

6.6.3 Model checking

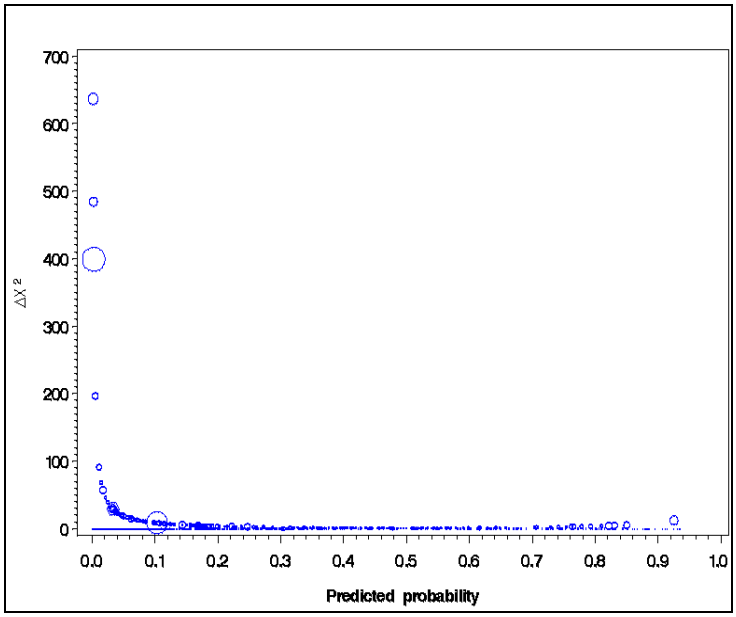
The validity of the reduced model was examined using the same approach as that in §6.5.2.

Assessment by deletion

There were four observations that were poorly predicted by the model, as measured by changes in the Pearson chi-square statistic (ΔX^2) upon deletion of the observation from the model (Figure 6.10). Inspection of these observations revealed that they were all infants of high gestational age (30 to 32 weeks), of appropriate birth weight (1485 to 1980g), and within

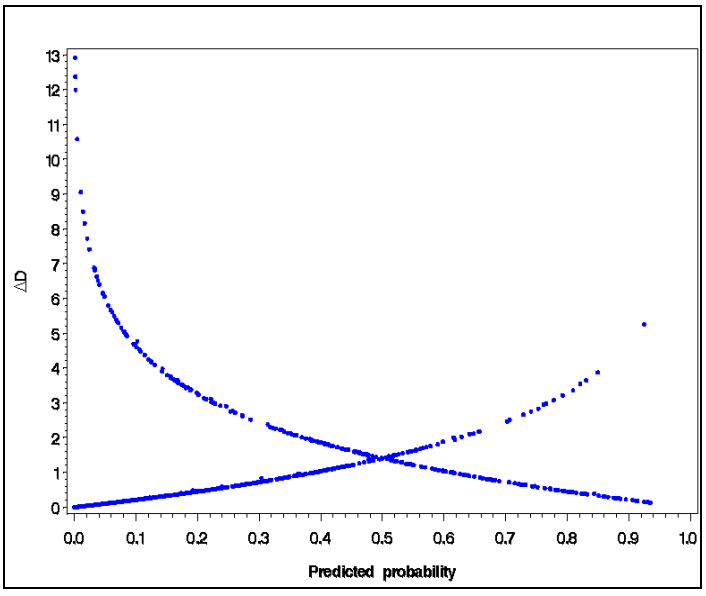
the range of base excess measurements considered to be normal, but who died before discharge from neonatal care. Each had a low predicted probability of death from the model (≤ 0.005) and, therefore, a large residual. These included the two observations noted with high values for ΔX^2 from the full model (Figure 6.1).

Figure 6.10 Change in Pearson chi-square statistic



The same four observations also had relatively high values for the change in deviance ΔD (Figure 6.11). Three of these observations were those found to have large values of ΔD from the full model (Figure 6.2). The values recorded for these observations were checked against the TNS forms and were found to be genuine observations.

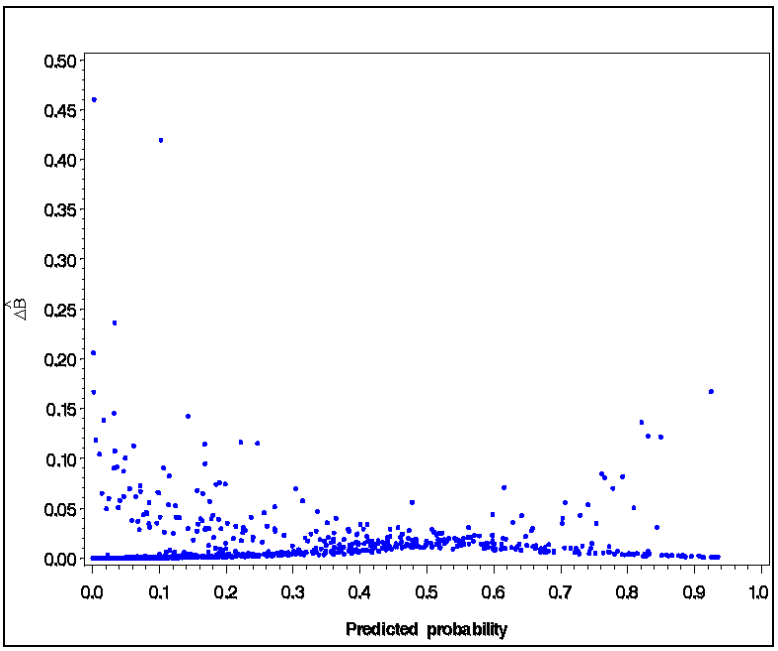
Figure 6.11 Change in deviance



All of the values of the DFBETAs were less than 1.0 (Appendix H), the minimum value suggested as signifying that an observation has a significant effect on the value of a parameter estimate (Hosmer & Lemeshow, 2000:180).

Once again, it was not the values of the parameter estimates themselves that were of primary importance, rather it was of more interest to consider changes in the vector of parameter estimates as a whole; $\Delta\hat{\boldsymbol{\beta}}_i$. In absolute terms the values were not large as all fell below 0.50. However, there were two observations with large values of $\Delta\hat{\boldsymbol{\beta}}_i$ relative to the other observations (Figure 6.12). One of these was one of the four observations identified in the previous plots. The other observation related to a relatively heavy (2250g) 32 week infant with a good Apgar score (9) but poor maximum base excess (-11.6mmol/L). Inspection of the individual DFBETAs for this observation (Appendix H) showed relatively high values for birth weight (-0.29) and the square of birth weight (0.33). While relatively extreme, none of the values were large in absolute terms.

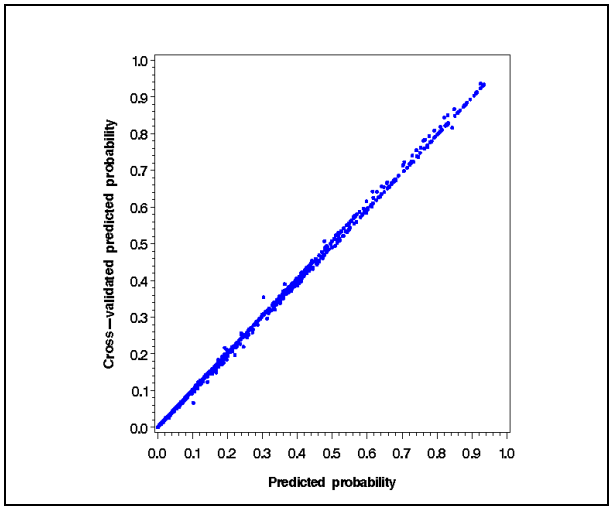
Figure 6.12 Change in model parameter estimate values



Internal model validation

The jacknife predicted probabilities were estimated using the one-step estimates outlined above (*6.2*) and are shown in Figure 6.13. There was good agreement between the predicted probabilities from the model and the jacknife estimated probabilities (Figure 6.13).

Figure 6.13 Cross-validated predicted probabilities



Calibration

The calibration of the model was investigated using the same approach as that described for the full model (§6.5). The calibration curve for all of the data is shown in Figure 6.14. There was good agreement between the observed and predicted mortality rates. Curves for two groups based on gestational age were also investigated (Figure 6.15).

Figure 6.14 Calibration curve: reduced model

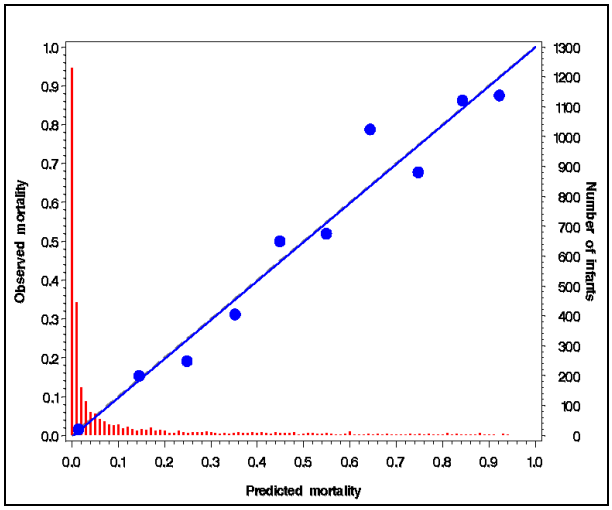
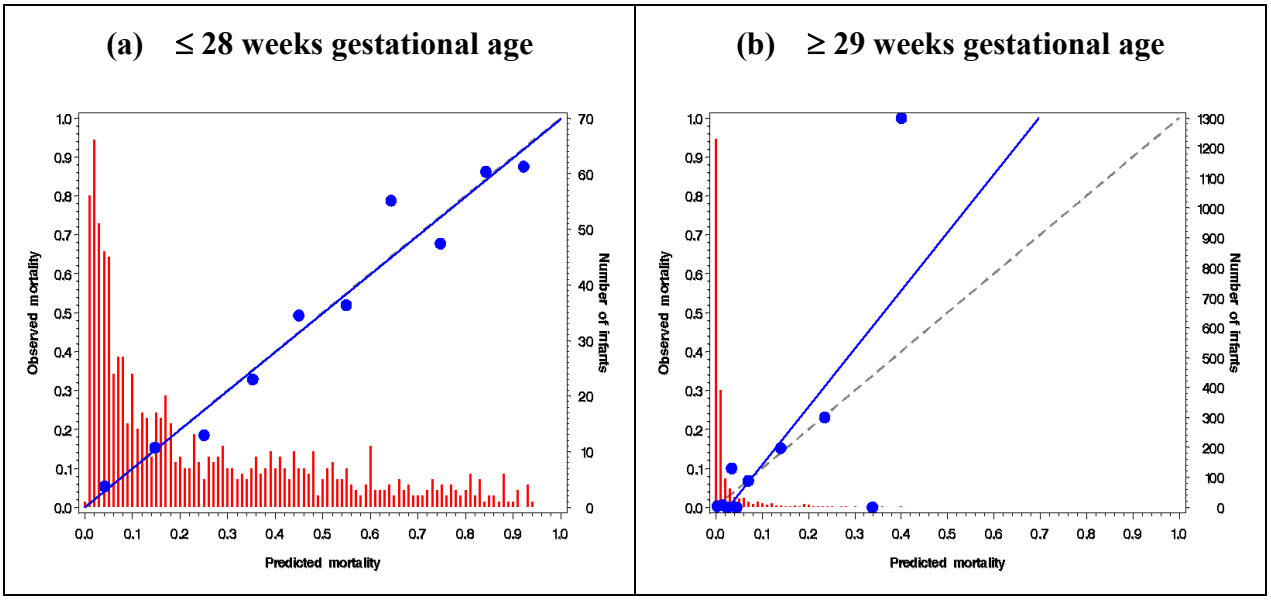


Figure 6.15 Calibration curves by gestational age group: reduced model

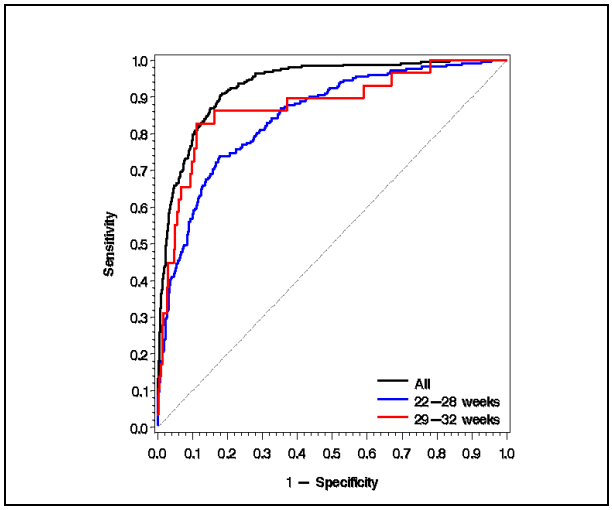


Once again the calibration of the model is questionable for those infants born at 29 weeks or over (Figure 6.15 b). However, there were only five observations with a predicted probability of death greater than 0.3, of which only one was greater than 0.4, so it was difficult to draw definitive conclusions from this plot.

Discrimination of the model

The ability of the model to discriminate between outcomes was investigated using the area under the Receiver Operator Characteristic (ROC) curve (§6.3.1). The value for the area under the ROC curve for all of the data was 0.932 (s.e. = 0.008), only slightly less that that estimated for the 'full' model (i.e. 0.940). When the observations were considered by gestational age, the values were 0.846 (s.e. = 0.015) for the 22 to 28 week infants and 0.878 (s.e. = 0.037) for the 29 to 32 week infants, both considered 'excellent' (Figure 6.16).

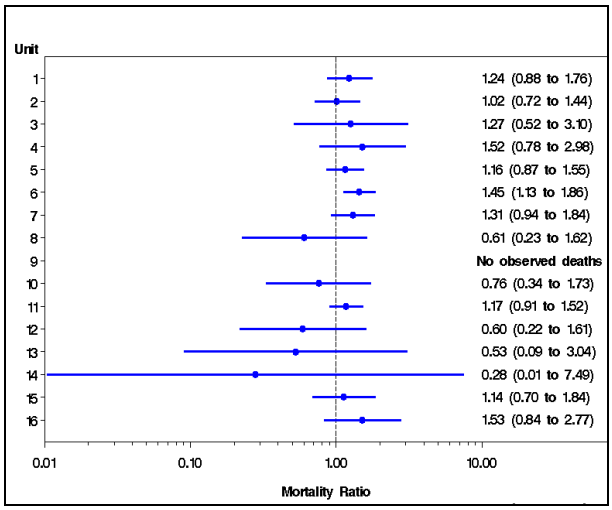
Figure 6.16 ROC curves: reduced model



Estimated Standardized Mortality Ratios

When SMRs were estimated using the reduced risk-adjustment model, the lower limit of the estimated 95% confidence interval for Unit 6 still took a value greater than one (Figure 6.17).

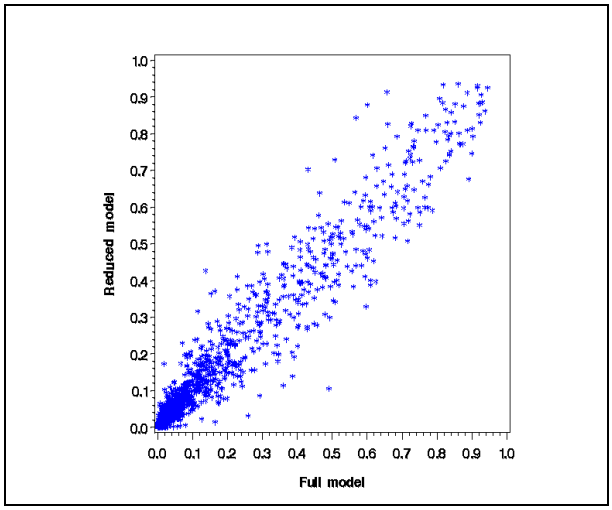
Figure 6.17 Estimated standardized mortality ratios: reduced model



6.7 Comparison of 'Full' and 'Reduced' Models

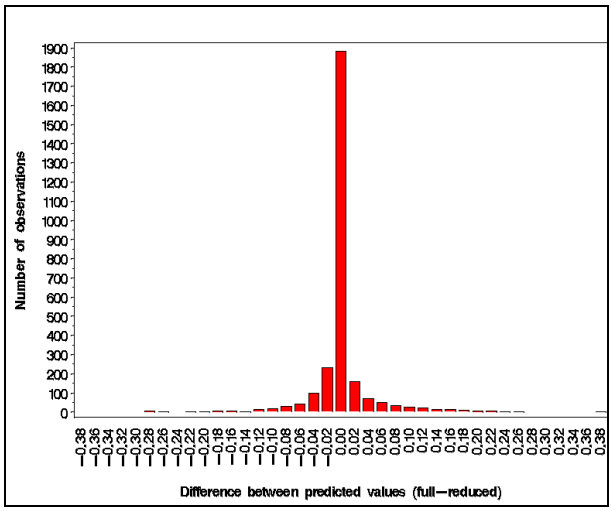
It is of interested to compare the predicted values from the two models presented in the previous Sections: 'Full' and 'Reduced' (Figure 6.18).

Figure 6.18 Predicted probability of mortality by risk-adjustment model



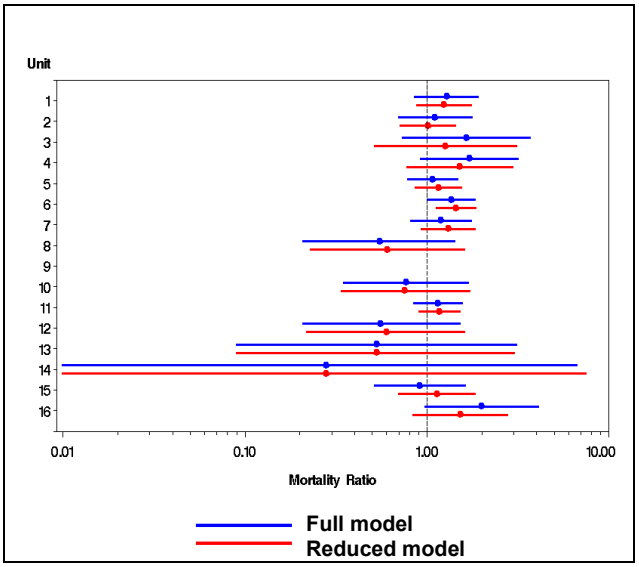
There were some differences between the predicted values from the two models. However, as was seen in Figure 6.5 and Figure 6.14, the majority of predicted values, for both models, were less than 0.1. If the individual differences are inspected (Figure 6.19), it can be seen that, although there are some large differences, most are less than 0.01. In addition, these differences seem to follow a symmetrical (and leptokurtic) distribution.

Figure 6.19 Differences in predicted mortality by model



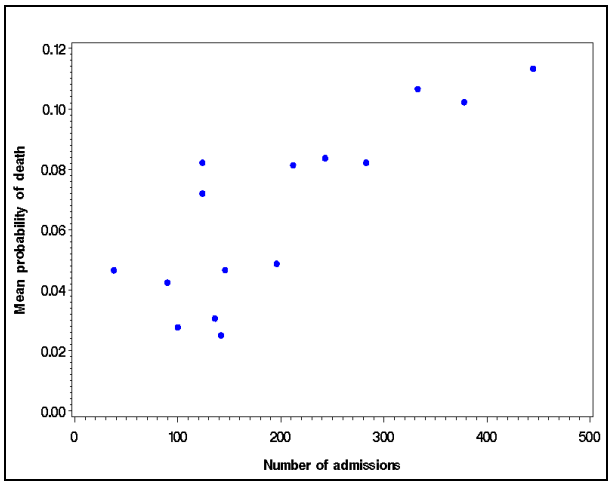
It was concluded that these differences were unlikely to alter any conclusions regarding the performance of an individual unit. Inspection of the estimated SMRs (Figure 6.9 and Figure 6.17) revealed minor differences between the estimates from the two models (Figure 6.20). In both cases, only Unit 6 did not have an estimated 95% confidence interval that contained the value one.

Figure 6.20 Estimated SMR: full model and reduced model



Although, in this case, the estimated SMRs were similar, there may be some evidence that important risk-adjustment variable were missing from the reduced model. The observed, unadjusted, mortality rates showed a trend of increasing rates with increasing unit size, as measured by the number of infants (Figure 4.1). This reflected the increased morbidity of infants in the larger units, as the sicker infants are often transferred from small units to large centres when addition facilities are available. This phenomenon is suggested by the increased mean predicted probability of death, as estimated by the full model, with increasing numbers of admissions (Figure 6.21).

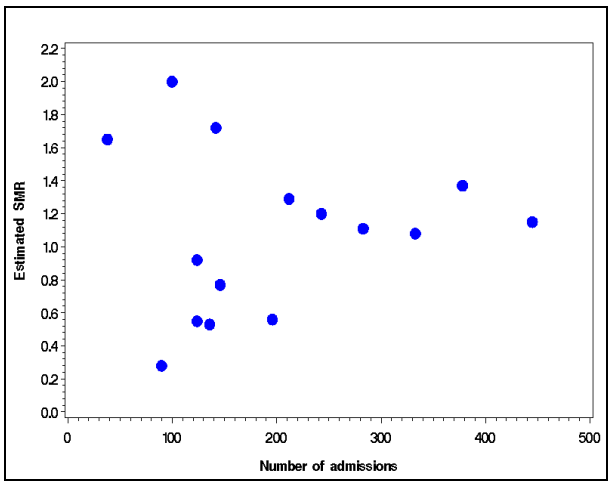
Figure 6.21 Mean predicted probability of death by number of admissions



If the risk-adjustment model accounted for all potential confounders the correlation between mortality rates and units size would no longer exist, if the underlying mortality rates were the same for all units. Under those circumstances a plot of SMRs against unit size would be

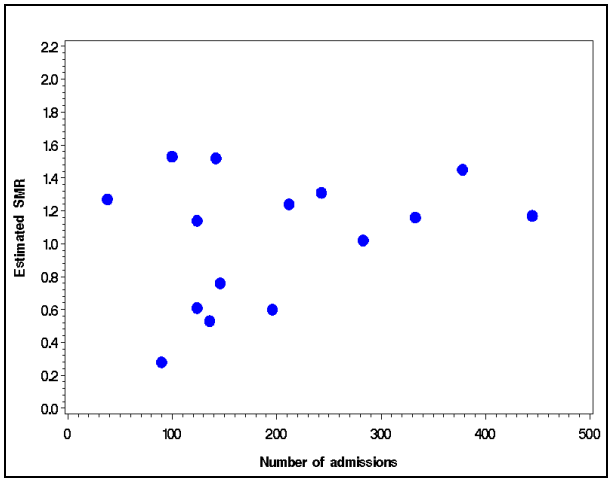
expected to show a 'funnel' pattern, with increased random variation associated with the small units. This can be seen when the estimated SMRs from the full model are plotted against unit size (Figure 6.22). This is a repeat of the methodology outlined in §3.3.2. As discussed there, once risk-adjusted outcomes are investigated the error around an estimate becomes a function of both the number of observation in a unit and of its case-mix. Therefore, smooth confidence limits (as shown in Figure 3.4) do not exist.

Figure 6.22 Estimated SMR by unit size: full model



However, a plot using the estimated SMRs from the reduced model showed some evidence of an excess of small units with low risk-adjusted mortality rates (Figure 6.23). This suggests that risk-adjustment was incomplete in this model and that clinically important variables were excluded from the reduced model. The relatively high mortality rates for the larger units shown in Figure 6.23 may result from genuinely higher rates in such units. However, the lack of evidence for such a pattern from the full risk-adjustment model (Figure 6.22) contradicts such a conclusion.

Figure 6.23 Estimated SMR by unit size: reduced model



6.8 Validation of 'Reduced' Model with 2003 Data

Although the aim of this thesis was to investigate births from 2000 to 2002, it was of interest to determine whether the reduced model was appropriate for subsequent cohorts of births. Trent Neonatal Survey data for births in 2003 became available during the writing of this thesis, and applying the model to these data may give an insight to its merit.

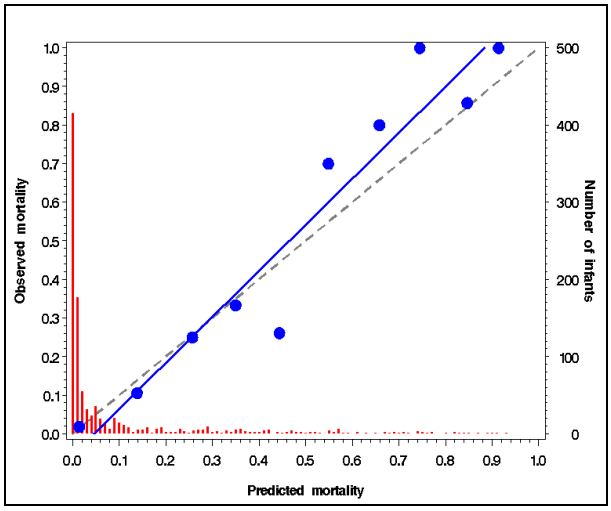
In 2003, there were 1108 infants admitted into Trent NICUs who met the inclusion criteria outlined in Chapter 2. One hundred and two of these infants (9.2%) died before discharge. However, 102 had missing data (101 for missing Apgar scores at one minute, and one for birth weight). This left 1006 infants of whom 91 died before discharge (9.0%).

The parameter estimates from the 'reduced' model (Table 6.3) were used to obtain an estimated predicted probability of death for each infant.

The area under the ROC curve indicated outstanding discrimination (A_{ROC} = 0.913: s.e. = 0.017).

The calibration of the 'reduced' model applied to TNS data from 2003 was investigated using a calibration plot (Figure 6.24) and the Hosmer & Lemeshow goodness-of-fit test ($\hat{C}$ = 24.23 ~ χ^2_8, p = 0.0021). The total number of predicted deaths was 86.9, underestimating the true number of deaths (91). The calibration plot suggested that the probability of death was underestimated for high-risk infants in particular.

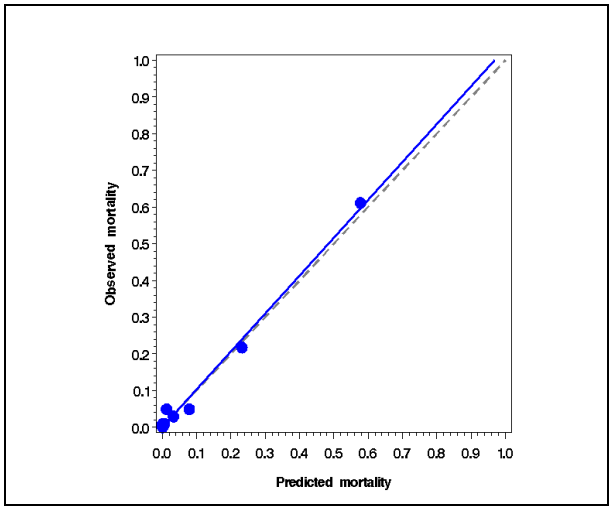
Figure 6.24 Calibration curve for 'reduced' model applied to 2003 data



However, a plot such as Figure 6.24 is disproportionately influenced by a relatively small number of observations with high predicted probabilities. An alternative plot can be created

using deciles of predicted risk, identical to those defined by the Hosmer and Lemeshow goodness-of-fit test (§6.3.2). In this case, each point represented approximately equal numbers of observations (Figure 6.25). There was still slight evidence that the model underestimated the risk of mortality for high-risk infants.

Figure 6.25 Calibration plot using deciles of risk



Further information on the calibration of the model could be obtained from Cox's modelling approach as the model was derived using different data from the validation data set (§6.3.3). The estimated model parameters shown no statistical evidence of poor calibration: $\hat{\alpha} = 0.01$ (s.e. 0.17), $\hat{\beta} = 0.95$ (s.e. 0.08). This was confirmed by the likelihood ratio tests (Table 6.4).

Table 6.4 Likelihood ratio tests for calibration and refinement

Null hypothesis	Test	Chi-square	d.f.	p-value
H_0: $\alpha = 0, \beta = 1$	$L(0,1) - L(\alpha,\beta)$	0.71	2	0.70
H_0: $\alpha = 0 \mid \beta = 1$	$L(0,1) - L(\alpha,1)$	0.33	1	0.57
H_0: $\beta = 1 \mid \alpha$	$L(\alpha,1) - L(\alpha,\beta)$	0.38	1	0.54

Although the likelihood ratio tests showed no evidence of poor calibration, the plots shown above suggested that re-estimation of the parameter estimates may be appropriate (Ivanov et al. 1999). Therefore, the model parameters were re-estimated using the 2003 TNS data (Table 6.5).

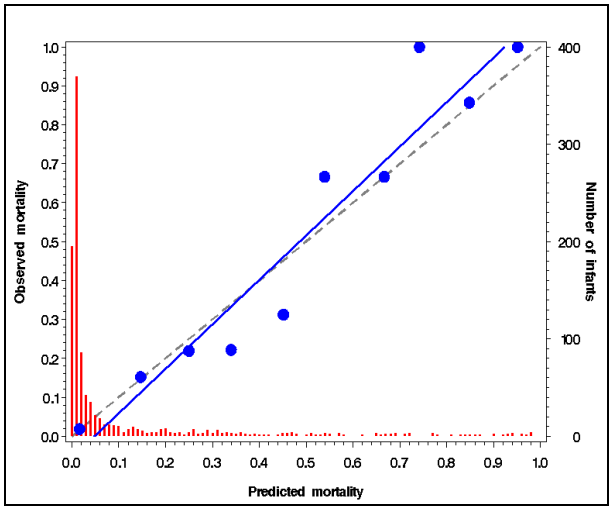
Most values for the parameter estimates were little different to those from the 2000-2002 data (Table 6.3). The largest differences were associated with low values of recorded base excess. Due to the smaller sample size, i.e. only one year's data, the standard errors were larger for the estimates based on the observations from 2003, leading to larger p-values.

Table 6.5 Parameter estimates for 'reduced' model from 2003 data

Variable	Group	$\hat{\beta}$	s.e.	P-value
Intercept		12.53	2.17	
Gestational age (week)		-0.31	0.09	0.0008
Sex	Female	Reference		0.97
	Male	0.01	0.30	
Birth weight (g)	Linear	-0.0082	0.0017	
	Quadratic	0.0000025	0.0000005	< 0.0001
Apgar		-0.17	0.07	0.010
Base excess	> -7.0	Reference		
	-7.0 to –9.9	-0.57	0.85	
	-10.0 to –14.9	-2.04	1.84	
	≤ -15.0	-2.48	1.50	
Base excess * birth weight	> -7.0 (mmol/L)	Reference		0.79
	-7.0 to –9.9	0.0004	0.0009	
	-10.0 to –14.9	0.0019	0.0023	
	≤ -15.0	0.0004	0.0013	
($A_{ROC} = 0.919$: $\hat{C} = 11.66 \sim \chi^2_8$, p = 0.17)				

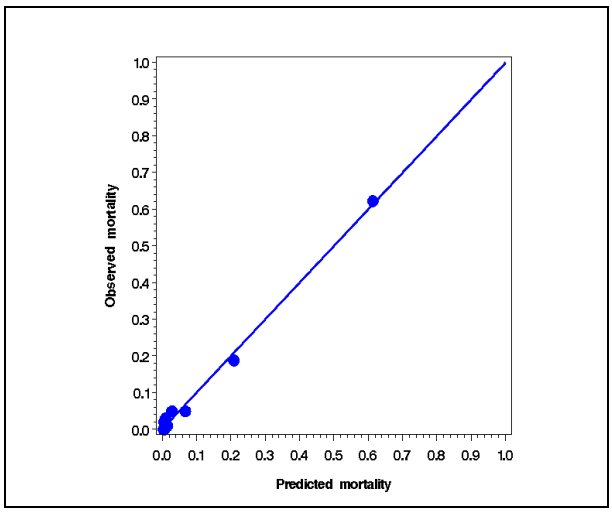
The value of the area under the ROC-curve was high (A_{ROC} = 0.919: s.e. = 0.016) and the Hosmer & Lemeshow goodness-of-fit test showed no evidence of poor calibration (p = 0.17). The calibration curve, based on ten groups defined the value of predicted probability, showed slight evidence of poor calibration (Figure 6.26).

Figure 6.26 Calibration curve for recalibrated 2003 data



However, once again each point represents a different number of infants and a small number of observations can influence the plot. When the curve was redrawn using the cut-offs from the Hosmer & Lemeshow goodness-of-fit test the calibration appeared to be excellent (Figure 6.27). Therefore, once recalibrated the model described the 2003 TNS data well.

Figure 6.27 Calibration curve for recalibrated 2003 data using deciles of risk



6.9 Sensitivity Analyses

Decisions have been made in the modelling process that may influence the results of the analysis. First, in §5.3.1 the use of the 'deviation from the mean' parameterization for the neonatal units was proposed to reduce the influence of the larger units. Second, infants without recorded Apgar scores at one minute were excluded from the analyses. Also, a stepwise model selection procedure was used to obtain the reduced model. These choices are explored in this Section. The probabilities estimated by the reduced model are also compared to the probabilities obtained using CRIB and CRIB II as the risk-adjustment methods.

6.9.1 Choice of parameterisation for the reference units

The 'deviation from the mean parameterization' was chosen in order to reduce the influence of the larger unit by comparing the outcomes in the unit of interest with the mean outcome of the reference units. This parameterisation gives equal weights to each unit rather than each infant. While the reference log odds were given at the mean of the log odds of the reference units, there was concern that the log odds for each reference unit are not estimated with equal precision. In particular, the very small units may have associated log odds with very large

variances. Such poor estimates will affect the estimated mean log odds, particularly if there are a small number of reference units.

In order to assess any bias introduced by poorly estimated unit effects, the 'reduced' model was repeated using the 'rest of Region' parameterisation described in §5.3.1. The model to estimate the model parameters for the reference data $\boldsymbol{\beta}_R$ was, where Unit j is the unit of interest:

$$log_e\left(\frac{\pi_{Ri}}{1-\pi_{Ri}}\right) = \beta_{R0} + \mathbf{X}_{Ri}\boldsymbol{\beta}_{Ri} \qquad (6.5)$$

where: $\mathbf{X}_R$ is the design matrix for the risk-adjustment variables

$\boldsymbol{\beta}_R$ is the vector of parameter values for the risk-adjustment variables

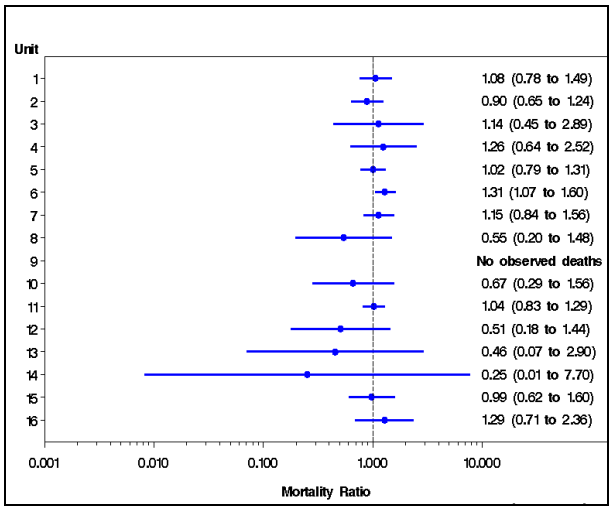
Hence, applying the parameter estimates to the observation from Unit j gives the estimated probability of death for an individual in unit j is:

$$\hat{\pi}_i = \frac{1}{1+e^{-\left(\hat{\beta}_0+\hat{\boldsymbol{\beta}}.\mathbf{x}_i\right)}} \qquad (6.6)$$

Summing these gives the expected total deaths in Unit j: i.e. $\sum_{i=1}^{n_j}\hat{\pi}_i$.

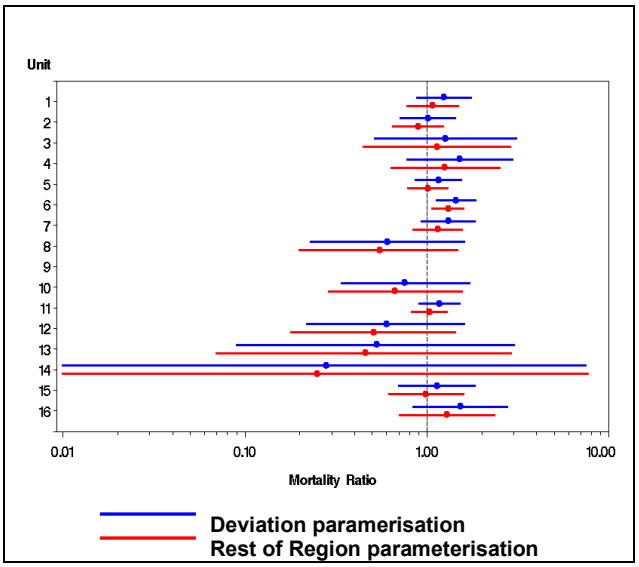
Since (*6.4*) has no term for the reference units, all of the reference observations are given equal weight. This model was applied to the TNS data, using the risk-adjustment variables from the 'reduced' model Figure 6.28.

Figure 6.28 Estimated SMRs from reduced model using 'rest of Region' parameterisation



When compared to the reduced model with deviation contrasts (Figure 6.17), this change in parameterization produced lower point estimates for the SMRs and lower values for the limits of the confidence intervals (Figure 6.30).

Figure 6.29 Estimated SMR: deviation and 'rest of Region' parameterisation



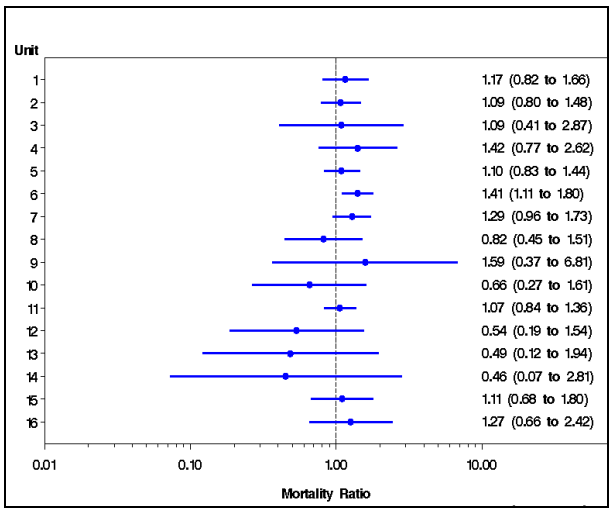
This behaviour is the same as that seen in §5.3.1. The change in parameterisation did not change the conclusions from the analysis: there was still evidence that Unit 6 has a SMR greater than unity.

6.9.2 Missing Apgar scores

The second modelling decision investigated was the inclusion of Apgar score at one minute in the model. The primary reason for investigating the sensitivity of the model to the inclusion of this variable was that 139 (4.6%) of the infants did not have a value recorded, of whom 34 (24.5%) died. However, there are two other reasons to investigate this. First, since the only death recorded for Unit 9 had a missing Apgar score, no SMR has been estimated for this unit. Second, the use of Apgar score may be criticized as it may potentially suffer from poor inter-rater reliability (Letko, 1996; Livingston, 1990) or be open to deliberate manipulation.

The reduced model was repeated without the inclusion of Apgar score at one minute and the SMRs estimated (Figure 6.30). The exclusion of Apgar score did not alter the conclusions of the analyses. There were only small changes to the point estimates and the confidence limits, and Unit 6 remained the only unit to have a lower limit for the 95% confidence interval above the value one.

Figure 6.30 Estimated Standardized Mortality Ratios: reduced model not including Apgar score at one minute



6.9.3 Model selection procedure

The model selection procedure for the reduced model developed in §6.6 was forward stepwise. However, it is known that different selection methods can lead to different models (Bland, 1995). To investigate the sensitivity of the data to the model selection procedure two further approaches were investigated: **forward selection** and **backward elimination**.

Forward selection

The forward selection method was similar to the forward stepwise method carried out in §6.6. However, in this case terms were not excluded from the model at subsequent steps once they had been included in the model. As before, up to three-way interaction terms were considered for inclusion and the significance level for inclusion was set at 10%.

The final model selected using this procedure was the same as the reduced model obtained using the forward stepwise method (Table 6.5).

Backward elimination

For the backward elimination approach all potential terms and interactions were first included in the model. Terms were then removed from the model in order of their statistical significance. One removed a term could not be reintroduced into the model. The process stopped once only statistically significant terms remained in the model. The significance level for elimination was set at 10%.

The model selected contained more terms (Table 6.6) than that obtained using the forward selection methods.

Table 6.6 Model parameter estimates: Backwards selection

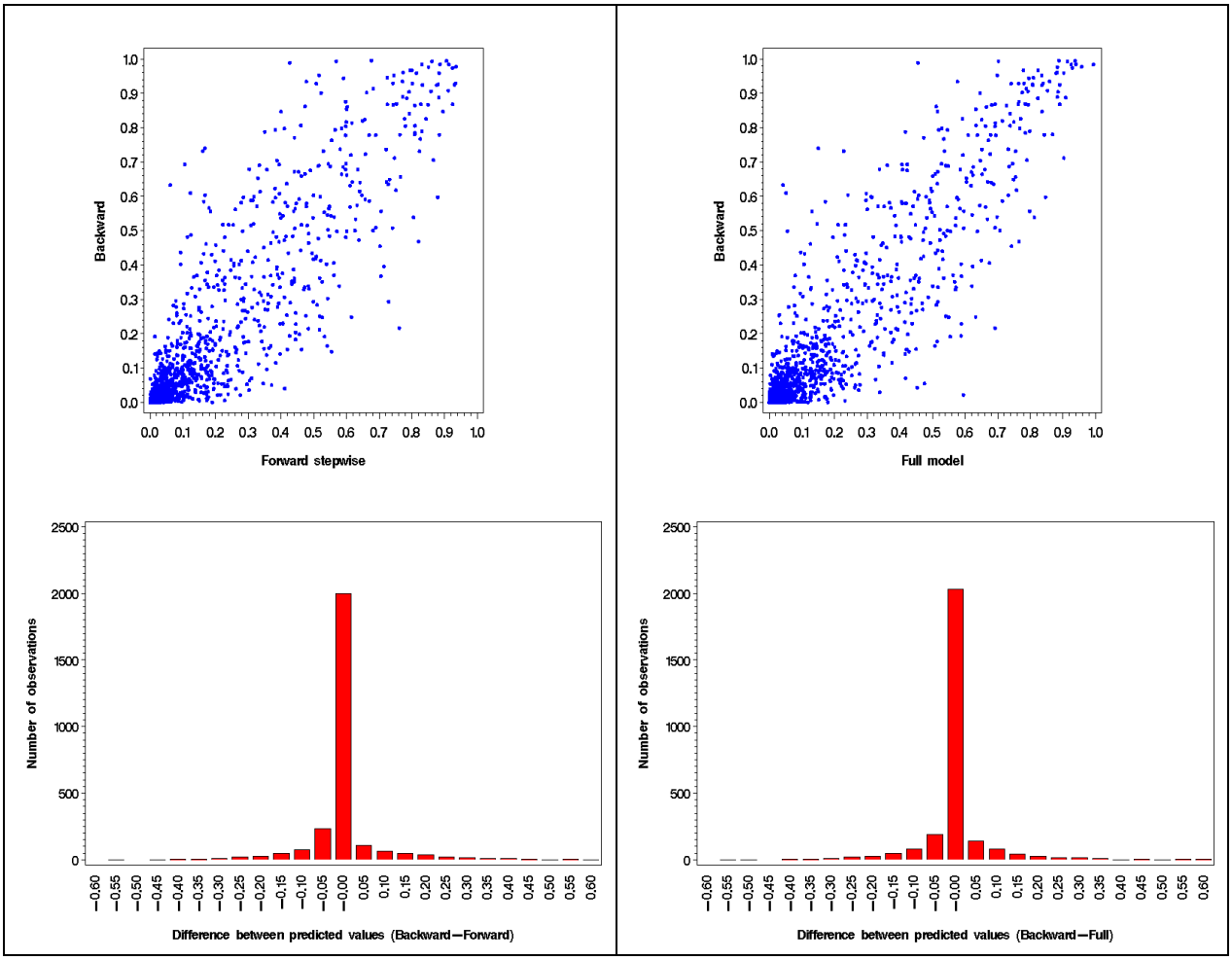
Variable	Group	$\hat{\beta}$	s.e.	P-value
Intercept		19.56	2.88	
Gestational age (week)		-0.48	0.11	
Sex	Female	Reference		
	Male	-1.33	0.45	
Birth weight (g)	Linear	-0.011	0.0015	
	Quadratic	0.0000029	0.0000005	< 0.0001
Apgar at 1 minute		-0.076	0.095	
Ethnicity	European	Reference		
	South Asian	-0.61	0.61	
	Other/unknown	0.42	0.58	
Congenital malformation	None	Reference		
	Present	-2.01	1.07	
Congenial malformation * ethnicity	Present*European	Reference		0.037
	Present*South Asian	0.41	1.12	
	Present*Other/unknown	3.22	1.25	
Base excess	> -7.0 (mmol/L)	Reference		
	-7.0 to –9.9	0.50	1.04	
	-10.0 to –14.9	-1.94	0.88	
	≤ -15.0	0.56	1.27	
Base excess * Birth weight	> -7.0 (mmol/L)* BWT	Reference		0.082
	-7.0 to –9.9 * BWT	0.00073	0.00096	
	-10.0 to –14.9 * BWT	0.0019	0.00073	
	≤ -15.0 * BWT	0.0013	0.0011	
Multiple birth	Singleton	Reference		
	Multiple	7.44	4.48	
Multiple birth * Gestational age	Singleton * gestation	Reference		0.013
	Multiple * gestation	-0.44	0.18	
Multiple birth * Base excess	Multiple * > -7.0	Reference		0.026
	Multiple * -7.0 to –9.9	0.17	0.73	
	Multiple*-10.0 to –14.9	2.18	0.76	
	Multiple * ≤ -15.0	1.46	1.03	
IMD		0.00014	0.017	
IMD * Sex	IMD * Female	Reference		0.0002
	IMD * Male	0.042	0.011	
IMD * Apgar at 1 minute		-0.0042	0.0024	0.073
IMD * Base excess	IMD * > -7.0	Reference		0.063
	IMD * -7.0 to –9.9	-0.030	0.015	
	IMD * -10.0 to –14.9	0.0058	0.015	
	IMD * ≤ -15.0	-0.031	0.019	
Corticosteroids	No	Reference		
	Yes	-7.11	2.74	
Corticosteroids * Gestational age	No	Reference		0.013
	Yes	0.25	0.10	
Fetal distress	No	Reference		
	Yes	-1.55	0.50	
Fetal distress * Sex	No * Male	Reference		0.072
	Yes * Male	0.71	0.39	
Fetal distress * Apgar at 1 minute	No * APGAR	Reference		0.0073
	Yes *APGAR	0.22	0.081	

Fetal distress * Congenital malform.	Yes * None	Reference		0.078
	Yes * Present	1.92	1.09	
Mode of delivery	Vaginal	Reference		
	CS: labouring	0.48	0.73	
	CS: non-labour	-0.59	0.51	
Mode of delivery * Base excess	CS:labour* > -7.0	Reference		0.0094
	CS:labour* -7.0 to –9.9	0.88	0.82	
	CS:labour*-10.0 to–14.9	1.42	0.88	
	CS:labour* ≤ -15.0	1.95	1.04	
	CS:non-lab* > -7.0	Reference		
	CS:non-lab* -7.0 to–9.9	-0.16	0.56	
	CS:non-lab*-10.0 to–14.9	1.65	0.58	
	CS:non-lab* ≤ -15.0	2.56	0.93	
Mode * Multiple birth	Vaginal * Multiple	Reference		0.032
	CS: labouring * Multiple	-1.70	0.99	
	CS: non-labour * Multiple	1.10	0.73	
Mother's age (year)		-0.011	0.018	
Mother's age * Multiple birth	Age * Single	Reference		0.035
	Age * Multiple	0.11	0.05	
Gravidity	Prima	Reference		
	Secund	-0.59	0.43	
	Terce	0.53	0.36	
Gravidity * Ethnicity	Secund * European	Reference		0.049
	Secund * South Asian	1.11	0.87	
	Secund * Other/unknown	0.21	0.85	
	Terce * European	Reference		
	Terce * South Asian	-0.46	0.76	
	Terce * Other/unknown	-1.70	0.78	
Gravidity * Mode of delivery	Secund * Vaginal	Reference		0.043
	Secund * CS: labouring	-0.59	0.84	
	Secund * CS: non-labour	0.12	0.60	
	Terce * Vaginal	Reference		
	Terce * CS: labouring	-1.40	0.68	
	Terce * CS: non-labour	0.72	0.49	
Infection	No	Reference		
	Yes	-0.18	0.26	
Infection * Multiple birth	Yes * Singleton	Reference		0.020
	Yes * Multiple	1.55	0.67	

Although the model showed good discrimination (A_{ROC} = 0.954), there was evidence of poor calibration ($\hat{C}$ = 34.88 ~ χ^2_8, p < 0.0001). However, the value for the Akaike Information Criterion (AIC) was 888.99, compared to 971.38 for the model selected using forward stepwise and 930.81 for the full model.

The predicted individual probabilities were compared to those from the forward stepwise model and those from the full model (Figure 6.31). Although differences were noted most were small in value.

Figure 6.31 Predicted probability of mortality by risk-adjustment model



However, the complexity of the model derived through backward elimination suggested statistical over-fitting of the data. Although this model showed excellent discrimination, as measured by the area under the ROC-curve (A_{ROC} = 0.954), the improvement was slight compared to the full model (A_{ROC} = 0.940) and the model derived using forward selection methods (A_{ROC} = 0.932). In addition, there was some evidence of poor calibration, not seen for the other two models. This model was also more complex than the alternatives. Therefore, it appears to offer no advantage over either of the other models.

6.9.4 Comparison with CRIB & CRIB II

The neonatal mortality risk-adjustment method used most often within the UK is CRIB, and its update CRIB II (§4.4.1). The approach taken in this thesis to develop a risk-adjustment model, rather than to use a pre-existing scoring system such as CRIB, was defended in §4.6. However, as described previously, TNS does collect data to allow the calculation of CRIB and, apart from temperature at admission, also CRIB II.

Temperature was not available here for two reasons: first, it had not previously been recorded by TNS (although collection was started in January 2004) and, second, it was felt that temperature at admission could be influenced by early neonatal care (§4.3). Hence, the CRIB II risk-adjustment variables available were gestational age, birth weight, sex and base excess, with interactions between the first three of these variables included in the score. In the model developed in Chapter 6 these same four variables wrere included, together with Apgar score at 1 minute and an interaction between birth weight and base excess.

The inclusion of Apgar score may be controversial, as it is open to manipulation, either intentional or unintentional. On the other hand, Apgar score has been shown to be associated with mortality (§4.4.8). Its advantage in this setting is that it is not based on gestational age or weight at birth and, therefore, may be able to grade the morbidity of infants of the same gestational age and weight. While much of the mortality rates can be 'explained' by these two important variables, it is the 'fine-tuning' of risk adjustment scores that is difficult. However, in this case, it was shown that excluding this variable from the model did not alter the final conclusions (§6.9).

To investigate any similarities between these risk-adjustment methods, two logistic regression models were estimated using CRIB and CRIB II as single explanatory variables and the predicted probabilities of death ($\hat{\pi}_i$) were estimated.

In general, there was strong correlation between the predicted probabilities of death using the three approaches (Table 6.7).

Table 6.7 Correlation of predicted values by risk-adjustment method

Risk-adjustment	**Spearman's rank correlation coefficient (ρ)**
CRIB & CRIB II	0.80
CRIB & Model	0.85
CRIB II & Model	0.94

The similarity between the estimates from the risk-adjustment model developed in this thesis and those from CRIB II was not surprising as, as discussed above, the reduced model derived in this thesis (§6.6) is similar to CRIB II. It should also be noted that for CRIB II the relationship between gestational age, birth weight and sex was derived using TNS data (Draper et al. 1999).

Figure 6.32 Comparison of predicted probabilities using CRIB and CRIB II

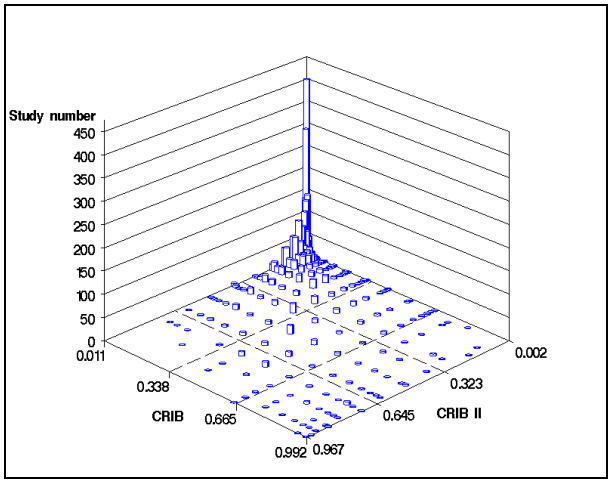
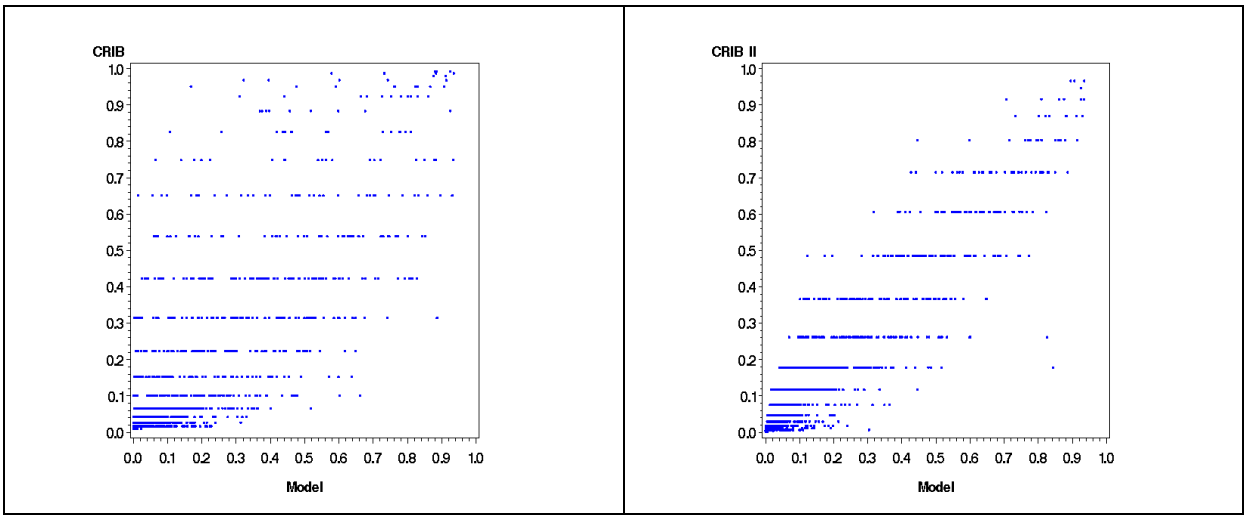


Figure 6.33 Predicted probabilities using CRIB and CRIB II compared to reduced model



Both the model derived using the data and the model with CRIB II showed no evidence of poor calibration, as measured by the Hosmer & Lemeshow Goodness-of-fit test (Table 6.8). All three models demonstrated excellent discrimination, with all three having areas under the ROC-curve of over 0.90. However, the model derived in this thesis had the lowest value for the AIC.

Table 6.8 Model fit statistics using different risk-adjustment methods

	A_{ROC}	H-L goodness of fit test	AIC
CRIB	0.923	$11.2770 \sim \chi^2_4$: p = 0.024	1091.268
CRIB II	0.914	$4.2240 \sim \chi^2_7$: p = 0.75	1134.689
Model	0.932	$6.4330 \sim \chi^2_8$: p = 0.60	971.384

The use of CRIB or CRIB II (without temperature at admission) would not have changed the overall conclusions of the analysis, i.e. that Unit 6 had statistically extreme outcomes, but the estimates SMRs would have differed (Figure 6.34 & Figure 6.35).

Figure 6.34 Estimated standardized mortality ratios adjusted for CRIB

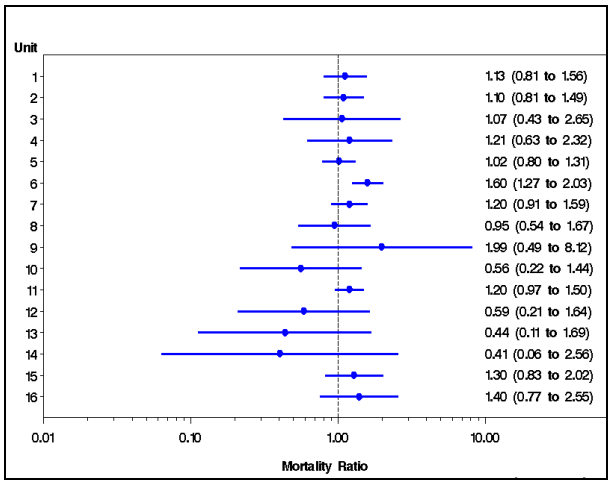
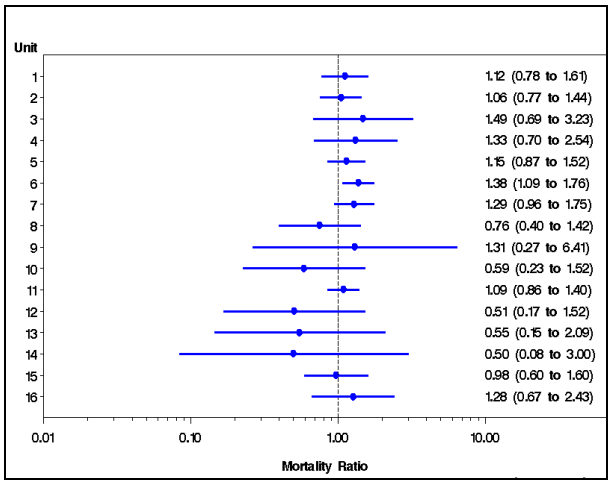


Figure 6.35 Estimated standardized mortality ratios adjusted for CRIB II



The use of CRIB or CRIB II would not have offered any advantage over the model developed in this Chapter, except perhaps for ease of use. On the other hand, the development of models in this Chapter has ensured that risk-adjustment has fully taken into account neonatal outcomes within the Trent Region.

In fact, the similarity between CRIB II and the reduced model lends confidence to both methods.

6.10 Bayesian Analyses

To illustrate a Bayesian approach, the reduced model was re-estimated to obtain the parameter estimates using Gibbs sampling, specifically using the WinBUGS software (§3.2.2) (Spiegelhalter et al. 1999b). All of the data were included, where possible, and vague prior distributions were specified for all hyper-parameters: *Normal*(0, 1000^2). Five chains, with diverse starting values, were inspected over a 1,000 iteration 'burn-in' was used to ensure sampling from the correct target distributions (Figure I.1). Once this had been confirmed a further 10,000 iterations were sampled to provide the parameter estimates. The WinBUGS code used is shown in Appendix I.1.

Estimation of parameter estimates for all data

The parameter estimates from the Bayesian model are shown in Table 6.9. The equivalent estimates from the classical model are shown for comparison, but the estimated value of the intercept has changed from Table 6.3 as the Bayesian model used gestational age centred at 30 weeks, to reduce autocorrelation in the sampled values.

Table 6.9 Bayesian fixed effects parameter estimates

Variable	Group	SAS		WinBUGS	
		$\hat{\beta}$	s.e.	$\hat{\beta}$	s.e.
Intercept		-4.94	0.43	-5.01	0.43
Gestational age (week)		-0.28	0.06	-0.29	0.05
Sex	Female	Reference			
	Male	0.39	0.17	0.41	0.17
Birth weight (g)	Linear	-2.44	0.66	-2.54	0.64
	Quadratic	2.37	0.49	2.31	0.48
Apgar		-0.13	0.04	-0.13	0.04
Base excess	>-7.0 (mmol/L)	Reference			
	-7.0 to –9.9	0.88	0.61	0.83	0.63
	-10.0 to –14.9	2.62	0.47	2.63	0.47
	≤ -15.0	3.62	0.51	3.66	0.51
Base excess * birth weight	>-7.0 (mmol/L)	Reference			
	-7.0 to –9.9	0.95	0.86	0.89	0.89
	-10.0 to –14.9	2.63	0.68	2.62	0.68
	≤ -15.0	2.90	0.79	2.91	0.79

The estimates for the parameters were very similar to those obtained using the classical approach. Density, auto-correlation and trace plots for each parameter are shown in Appendix I.1. There was no evidence of poor mixing or inappropriately high auto-correlation for any of the parameter estimates.

Estimation of SMRs

The 'reduced' model was then used to obtain risk-adjusted SMRs for all of the units, through the model described in § 5.8.4 (*5.54*), where for Unit *j*:

$$\hat{SMR}_j = \frac{\sum_{i=1}^{n_j} \hat{d}_i}{\sum_{k=1}^{n_j} \hat{\pi}_i} \qquad (6.7)$$

where: $\hat{d}_i = \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_j \mathbf{X}_i)}}$

$$\hat{\pi}_i = \frac{1}{1+e^{-(\hat{\beta}_{R0} + \hat{\boldsymbol{\beta}}_R \mathbf{X}_i)}}$$

n_j is the number of observations in Unit *j*

β_{R0} is the estimate of the mean log odds of the reference units obtained from a logistic regression model with the reference data; see (*5.4*).

$\boldsymbol{\beta}_R$ is the vector of parameter estimates for the risk-adjustment variables obtained from a logistic regression model with the reference data

The parameter estimates $\hat{\boldsymbol{\beta}}_R$ are obtained using data from the other 15 units:

$$\hat{g}_i = \hat{\beta}_{R0} + \hat{\boldsymbol{\beta}}_R . \mathbf{X}_i + \sum_{\substack{k=1 \\ k \neq j}}^{16} \hat{\beta}_k . I_k$$

X is the design matrix for the risk-adjustment variables and the unit indicator variables I_k follow the deviation parameterization described in §5.3.1.

The prior distributions were specified for all of the model parameters: β_{R0}, $\boldsymbol{\beta}_R$, β_0, $\boldsymbol{\beta}_j$. As was shown in §5.8.4, the choice of distribution was important as some of the units had very few observations (Spiegelhalter et al. 1999a). The parameter β_{R0} represented the log odds of deaths, for the reference population, where the values of the risk-adjustment variables are zero. The observed values for gestational age and birth weight were centred (at 30 weeks and

1500g respectively) to reduce auto-correlation of the parameter estimates between iterations, and so β_{R0} represented the log odds of deaths for an infant born at 30 weeks and 1500g, and with all other variables taking the value zero. The previously estimated probability of death for a girl born at this weight and gestational age was 0.01 (Draper et al. 1999), giving log odds of deaths of $\log_e\left(\frac{0.01}{1-0.01}\right) \approx -4.6$. This value was selected as the mean of the prior distribution for β_{R0} but with a large variance: $\beta_{R0} \sim Normal(-4.6, 1000^2)$.

The vector $\boldsymbol{\beta}_R$ contains the parameters values for the risk-adjustment variables for the reference population on a log odds scale. In the modelling here, each component of $\boldsymbol{\beta}_R$ was given the prior distribution $Normal(0, 1000^2)$. The distribution $Normal(0, 1)$ was specified as the prior probability distribution for each of the parameter estimates (β_k) for the unit indicator variables, as discussed in §5.8.4.

The equivalent parameters for the unit of interest are given by β_0 and $\boldsymbol{\beta}_j$. The data available to estimate values for the parameters were sparse for all but a few units. Since, *a priori*, there was no reason to assume that the parameter values for the unit of interest were different from those of the reference population, the mean of the prior distributions for the unit of interest were specified as the value of the reference population estimates but with a smaller variance than the parameters of the reference units: e.g. $\beta_0 \sim (\hat{\beta}_{R0}, 10)$. Since the estimates were obtained in a single model, the value of $\hat{\beta}_{R0}$ used was the current value at each iteration. The *cut()* function was used in the WinBUGS code to avoid the potential for flowback, thus avoiding any possibility of the data from Unit *j* influencing the parameter estimates for the reference population.

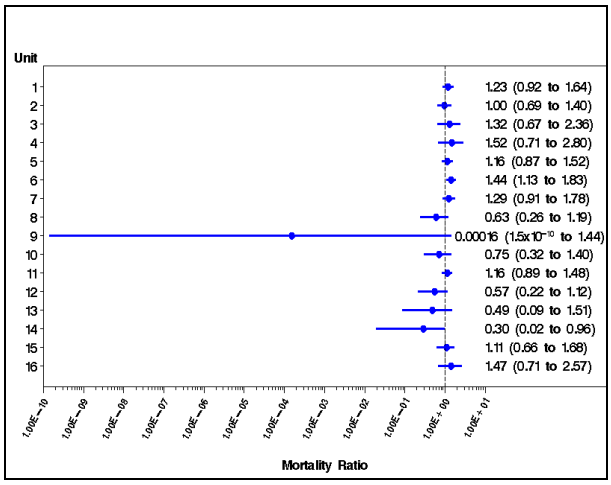
Alternative informative prior distributions could have been chosen derived from elicited beliefs, previous years' data, or an adaptation of 'sceptical priors' (Spiegelhalter et al. 1994; Parmar et al. 2001). However, the aim here was to illustrate the potential usefulness of a Bayesian approach, rather than to specifically investigate the influence of prior distribution on estimates obtained from this model.

The posterior probability that the value of the SMR was under ⅔, or over 3/2, was estimated by monitoring the proportion of sampled values that lay outside these limits. An example of the WinBUGS code used is given in Appendix I.2. There was no evidence in any of the models for poor mixing or high auto-correlation (plots not shown).

Table 6.10 Bayesian estimated SMRs: reduced model

Unit	Estimated SMR (95% CI)		P(SMR<2/3)	P(SMR>3/2)
1	1.23	(0.92 to 1.64)	0.0003	0.085
2	1.00	(0.69 to 1.40)	0.016	0.0073
3	1.32	(0.67 to 2.36)	0.023	0.33
4	1.52	(0.71 to 2.80)	0.018	0.52
5	1.16	(0.87 to 1.52)	0.0002	0.029
6	1.44	(1.13 to 1.83)	<0.0001	0.37
7	1.29	(0.91 to 1.78)	0.0001	0.18
8	0.63	(0.26 to 1.19)	0.56	0.0021
9	1.6×10^{-4}	(1.5×10^{-10}to1.44)	0.95	0.23
10	0.75	(0.32 to 1.40)	0.37	0.013
11	1.16	(0.89 to 1.48)	<0.0001	0.019
12	0.57	(0.22 to 1.12)	0.66	0.0017
13	0.49	(0.09 to 1.51)	0.70	0.014
14	0.30	(0.02 to 0.96)	0.90	0.0012
15	1.11	(0.66 to 1.68)	0.026	0.077
16	1.47	(0.71 to 2.57)	0.018	0.47

Figure 6.36 Estimated standardized mortality ratios: Bayesian reduced model

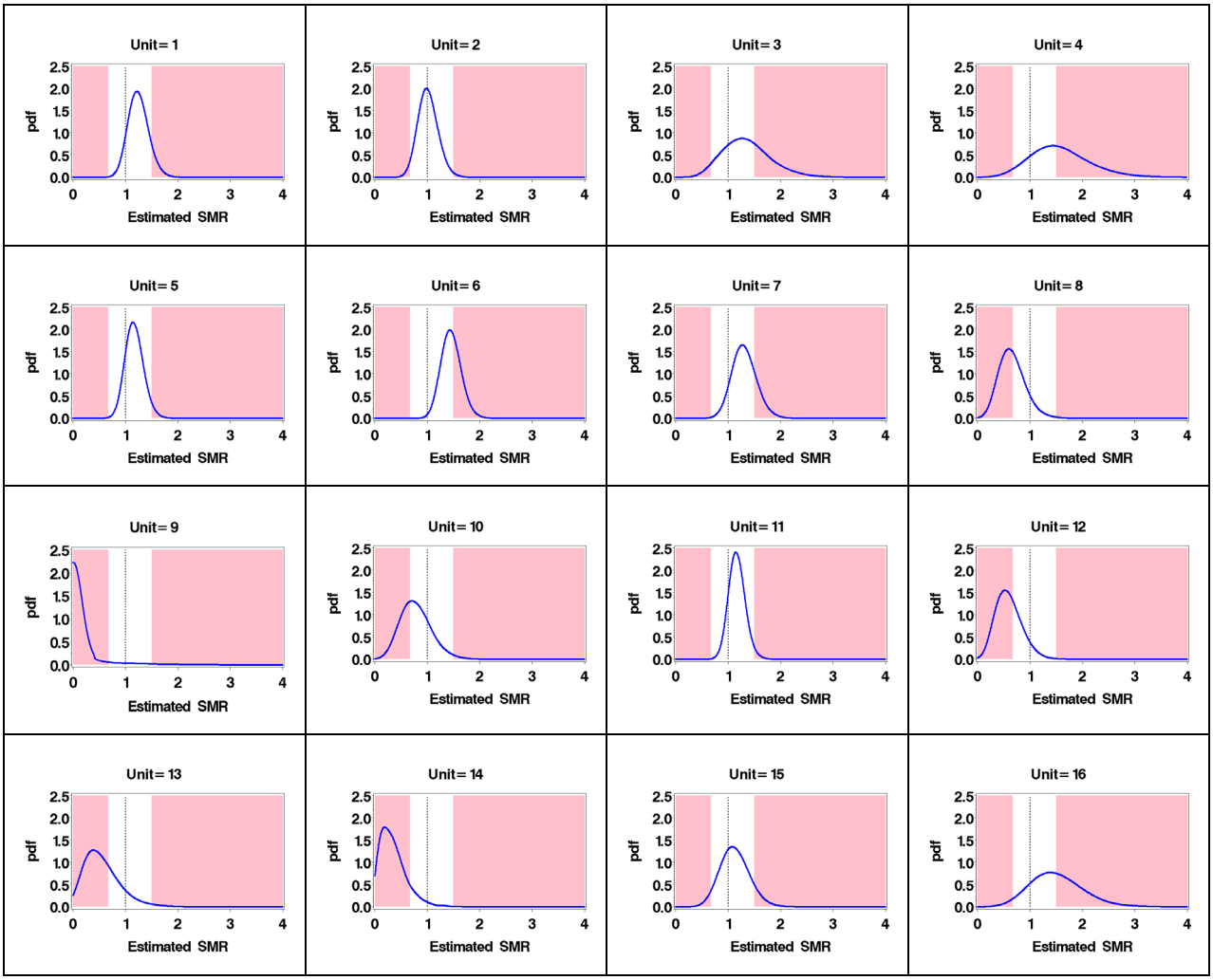


The values of the estimated SMRs (Table 6.10) were similar to those from the classical model (Table 6.3). These small differences made no qualitative difference to the conclusions that could be drawn from the model through the estimated 95% confidence intervals, except in two cases. In the Bayesian analysis the upper limit of the credible interval for Unit 14 fell just

below the value 1.00. It has previously been shown in this thesis that the method of Hosmer & Lemeshow tends to produce confidence intervals with high values for the upper limits (§5.6). This was especially true for small units, such as Unit 14. The second difference is that, using the Bayesian approach, it was possible to estimate a SMR, and credible interval, for Unit 9. This is discussed further in §6.11.

A further advantage of the Bayesian approach is that probability statements could be made about the SMR: the probability that the true value of the SMR for each unit lies below ⅔, or over ½ is shown in Table 6.10. Smoothed posterior density functions are shown in Figure 6.37. Unit 4 was more likely than not to have a true SMR of over ½: i.e. $P(\text{SMR} > \frac{1}{2}) > 0.5$, while Units 8, 9, 12, 13 and 14 showed similar evidence for low SMRs.

Figure 6.37 Smoothed posterior probablilty density functions for SMR

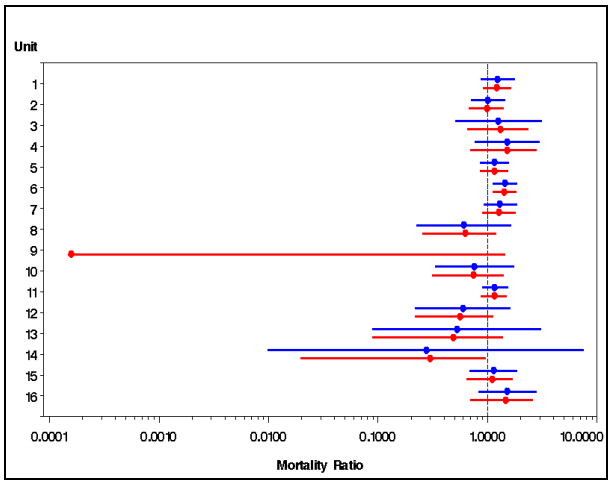


Comparison with frequentist approach

The estimation of posterior probability distributions allowed straightforward clinical interpretations to be placed on model parameters, rather than solely relying on statistical criteria. In both the frequentist and the Bayesian reduced models, Unit 6 had a 95% confidence (credible) interval that excluded the value one (Figure 6.38). In the Bayesian analysis the estimated credible interval for Unit 14 also did not contain the value one but this was not seen in the frequentist analysis. The tendancy for the Hosmer & Lemeshow method, of estimating confidence intervals for the SMR, to produce high upper limits has been noted previously (Figure 5.19).

However, the Bayesian model also identified Units 8, 9, 12, 13 and 14 (where $P(\text{SMR}<\frac{2}{3}) > 0.5$) and Unit 4 (where $P(\text{SMR}>\frac{1}{2}) > 0.5$) as units 'more likely than not' to have clinically extreme risk-adjusted mortality rates. Therefore, the Bayesian approach potentially allows small units with extreme rates, but large confidence intervals, to be identified for further investigation, as well as the identification of large units with statistically significant, but clinically unimportant, rates.

Figure 6.38 Frequentist and Bayesian estimated SMRs: reduced model



6.11 Units with no observed deaths

Unit 9 was excluded from the frequentist analyses estimating the SMRs, in §6.5, §6.6 because there were no observed deaths for this unit once observations with missing data were excluded. The absence of deaths caused problems both when the unit was part of the reference population and when it was the unit under investigation. In this case, imputation of

the missing Apgar value would have allowed the inclusion of all observations in the model (Zhang, 2003), and thus solved the problem of no events. However, in other cases there may be units that genuinely have no observed deaths (or, although more unlikely, no observed survivors) over the time period under investigation. This is particularly a problem when short time periods are used, the event is rare, or when the providers encounter small numbers of patients. However, the problems outlined below are mathematical. In principle, units with no observed deaths, and, indeed, with no observed survivors, should be included in the analyses as these are likely to be units with extreme performances.

Part of reference population

When using the deviation parameterization model set out in (*5.4*) (§5.3.1), the absence of observed deaths in a reference unit produces quasi-complete separation of the data. This results in unstable parameter estimates from the logistic regression model (§3.4). A potential solution is to use a different parameterization of the model that does not include parameters for the individual reference units, for example the 'rest of Region' and 'weighted' models illustrated in §5.3.1. An alternative approach is to use a random effects model (§5.10), which would not require the estimation of unit-specific model coefficients. However, potential disadvantages of both of these approaches have been discussed (§5.3.1 & §5.10) and it is unclear which method would be the most appropriate. In this thesis such units (i.e. Unit 9) were excluded from the reference population, but it is recognised that, using such an approach, the mortality rate in the reference population may be upwardly biased.

Unit under investigation

Investigating outcomes in units with no observed deaths, such as Unit 9, presents a problem as some of the methods of estimating confidence intervals for the SMR discussed in §5.6 cannot be used. The three methods based on the Normal approximation that only include the uncertainty from the observed number of deaths (i.e. 'with CC', 'without CC' and 'full') can be applied to such units, as the uncertainty is quantified by $\sum(\pi_i[1-\pi_i])$. This approach has been taken by both the New York State Department of Health (2004) and Papworth Hospital (Papworth, 2005) for the reporting of results for cardiac surgeons. However, it was shown in §5.7 that such methods perform poorly when the observed SMR is far from the null hypothesis (i.e. unity). When applied to Unit 9 the estimated 95% confidence intervals were:

'With CC' (-4.16 to 4.16)

'Without CC' (-2.90 to 2.90)

'Full' (0.24 to 10.80)

The first two methods produce intervals symmetrical about the point estimate for the SMR (i.e. zero), when negative values are implausible. The coverage properties of such intervals are unknown, but are likely to be very poor. The 'full' method generated a lower limit for the confidence interval that was greater than the point estimate (see also Appendix C.4). This is clearly inappropriate.

The other frequentist approaches set out in §5.6 would fail when the unit of interest has no observed events. Both of the bootstrap methods would only produce simulated values for the SMR of zero, since the sample numerator would always be zero. The extensions to the Normal approximation method proposed by Hosmer & Lemeshow (1995) and by Zhou & Romarno (1997) cannot be used as they require devision by the observed number of deaths; see (*5.45*) and (*5.53*).

The Bayesian model developed in §5.6.4 offered a solution. When applied to Unit 9 (using the conditions given in §6.10) the estimate for the SMR was 1.6×10^{-4} (95% credible interval: 1.5×10^{-10} to 1.44), with $P(\text{SMR} < \frac{2}{3}) = 0.95$ and $P(\text{SMR} > \frac{3}{2}) = 0.023$. However, it is unclear how sensitive these estimates are to the choice of prior distributions. Alternative, plausible, prior distributions for β_0 were investigated, keeping all other prior distributions the same as in §6.10. The estimates for the SMR and the reported posterior probabilities were all sensitive to the choice of prior distribution (Table 6.11).

Table 6.11 Estimated SMR for Unit 9 given various prior distributions

Prior distribution β_0	**SMR**	**(95% CI)**	**$P(\text{SMR} < \frac{2}{3})$**	**$P(\text{SMR} > \frac{3}{2})$**
$N(\beta_{R0}, 10)$	1.59×10^{-4}	$(1.54\text{x}10^{-10}$ to 1.44)	0.95	0.023
$N(0, 10)$	2.64×10^{-3}	$(1.68\text{x}10^{-8}$ to 1.90)	0.91	0.039
$N(0, 10^2)$	2.70×10^{-3}	$(1.80\text{x}10^{-8}$ to 1.91)	0.92	0.036
$U(-5, 5)$	5.50×10^{-4}	$(3.12\text{x}10^{-10}$ to 1.61)	0.93	0.029

An alternative Bayesian approach proposed only included the uncertainty from the expected number of death (Austin *et al*, 2001; Austin, 2002), that is, the numerator is the observed number of deaths. From (*6.6*) this is given by:

. $$\hat{SMR}_j = \frac{\sum_{i=1}^{n_j} d_i}{\sum_{k=1}^{n_j} \hat{\pi}_i}$$

where: $$d_i = \begin{cases} 0 & \text{if infant } i \text{ discharged alive} \\ 1 & \text{if infant } i \text{ died before discharged} \end{cases}$$

$$\hat{\pi}_i = \frac{1}{1 + e^{\left(-\beta_{R0} - \hat{\boldsymbol{\beta}}_{\mathbf{R}} \mathbf{X_i}\right)}}$$

n_j is the number of observations in Unit j

$\boldsymbol{\beta}_R$ is the vector of parameter estimates obtained from a logistic regression model with the reference data

β_{R0} is the estimate of the mean log odds of the reference units obtained from a logistic regression model with the reference data; see (*5.4*).

However, it was observed in §5.8.5 that uncertainty from the observed number of deaths (i.e. the numerator of the SMR) dominates the estimation of the variance for the SMR. Ignoring this source of uncertainty would result in estimated credible intervals being too narrow. Indeed, the application of this approach to the simulated datasets described in §5.7 resulted in 95% credible intervals with coverage rates for the various scenarios from 37% to 73%.

Further developments and extensions

The existence of units with no observed deaths presents particular problems to the modelling approach developed in this thesis. The exclusion of Unit 9 from the reference population in the final models is unlikely to have 'significantly' affected the results for the other units. The absence of an estimated SMR for Unit 9 is unfortunate, as this unit is then excluded from the Regional monitoring process. However, for such a small unit it is unlikely that there would be statistical evidence that its true SMR is not unity, as suggested when Apgar score was excluded from the model (§6.9). This may allow, with caution, the reporting of intervals based on the Normal approximation, such as the 'with CC' interval illustrated above. Should such a situation arise with a large unit, then the methodology used to estimate the confidence interval may become more important.

The Bayesian estimates reported above showed a probability of 0.95 that the true SMR was below ⅔, even though the 95% credible interval contained unity. If true, then this is an

important conclusion, although the result from this particular analysis must be treated with caution because of the effect of the choice of prior distributions.

It is clear that it is important that such units are not excluded from analyses as, potentially, there may be important lessons to be learnt from them.

6.12 Chapter Summary

In this Chapter variables recorded by TNS and thought to be associated with in-unit mortality were investigated. Estimated SMRs, with 95% confidence intervals, were obtained for each unit, where possible. A model including all of these variables was produced (§6.5), together with a reduced model with variables selected according to their statistical significance (§6.6). Both models showed good discrimination and calibration. There was evidence from these models that Unit 6 had an in-unit mortality rate greater than the units in the rest of the Region (§0). When recalibrated, the reduced model performed well with TNS data from 2003 (§6.8) and the conclusions held when an alternative parameterization of the reference model was used and when Apgar score at one minute was excluded from the model (§6.9).

In §6.10 a Bayesian model was developed and illustrated. This approach allowed the use of probability statements for the outcome statistics, in this case $P(\text{SMR}<\frac{2}{3})$ and $P(\text{SMR}>\frac{3}{2})$, in addition to the estimation of point estimates and credible intervals. Although Unit 6 was the only unit whose 95% credible interval did not contain unity, there was evidence that Units 8, 9, 12, 13 and 14 were more likely than not to have a SMR under $\frac{2}{3}$, while Unit 4 showed similar evidence for SMRs over $\frac{3}{2}$.

Finally, the problem of no observed deaths, encounted with Unit 9, was discussed and possible solutions outlined (§6.11).

Chapter 7: DISCUSSION & CONCLUSIONS

7.1 Chapter Overview

The structure and main findings of this thesis are described in §7.2. The clinical and statistical implications of this work are discussed in §7.3. In §7.4 the limitation of this study are described and identified potential further work is outlined in §7.5. Section 7.6 comprises the conclusions drawn from this work.

7.2 Summary

7.2.1 Overall structure

The aim of this thesis was to critically review, and where appropriate develop, statistical methods in order to identify the most appropriate methods for the estimation of summaries of in-unit mortality rates for neonatal intensive care units in the former Trent Heath Region. In particular, mortality rates for infants born at less than 33 weeks gestational age, in the years 2000 to 2001, admitted to each unit were investigated and compared to the combined mortality rate of the other units.

The three stages of the process of producing such profiles were described in Chapter 1: measurement, analysis and action. Chapter 2 comprised an introduction to the data analysed in the thesis and their source, the Trent Neonatal Survey (TNS). The unique nature of these data was highlighted, together with the need for their proper analysis. This Chapter also contained a brief description of neonatal care, its organization and issues relevant to neonatal mortality.

In Chapter 3 various statistical methods that may have a function in the production of provider profiles were described and illustrated. Simple methods that may be useful in preliminary data analyses were described and illustrated in §3.3. However, such methods are inadequate for a comprehensive investigation of in-unit mortality rates and it was suggested that logistic regression models provided a more flexible and robust approach. Other statistical methods that may be useful in provider profiling using binary measures, but were not appropriate for the data in this thesis, were also briefly described (§3.5).

While the statistical methods described in Chapter 3 quantify the random variation in the observed in-unit mortality rates, a unit may have an extreme rate due to the morbidity of the population admitted. The need to account for any differences in the population structures between the units that may cause differences in observed mortality rates was explained in Chapter 4. Published risk-adjustment scores were identified, appraised, and their use discussed. It was stated that none of these pre-existing scoring systems were suitable for the data in this thesis and that risk-adjustment would be through a specifically derived logistic regression model.

The selection of the most appropriate summary statistic was considered in Chapter 5. Odds ratios were illustrated first. At the same time, three approaches to the parameterisation of the reference units were explored and the use of deviation contrasts in subsequent analyses was justified. Difficulties exist in the interpretation of odds ratios, and summary measures derived from standardization have a clear clinical interpretation. For this reason, direct and indirect standardization were described and compared in §5.4. Mortality ratios using directly standardized outcomes (Comparative Mortality Figure) and indirectly standardized outcomes (Standardized Mortality Ratio) were described and illustrated using the TNS data. The use of the SMR for this thesis was argued. The properties of various methods proposed for estimating confidence intervals for the SMR, together with a Bayesian approach proposed here, were investigated through a simulation study and the use of the method from Hosmer and Lemeshow (1995) was proposed for this thesis. Random-effects models were also illustrated and discussed.

Chapter 6 comprised an investigation into the relationship between in-unit mortality and infant, perinatal and antenatal factors. Each variable recorded by TNS and thought, *a priori*, to be associated with mortality was investigated separately and then with other variables thought to interact with it, usually gestational age. These variables, and proposed interactions, were then included in a single logistic regression model and SMRs, with 95% confidence intervals, for the units estimated. A reduced model was also estimated using a forward selection method. Also in Chapter 6, model checking and validation methods were described and applied to the models, and a Bayesian approach to estimating the model parameters was developed.

7.2.2 Final modelling approach

The main analysis of the thesis employed logistic regression modelling to estimate standardized mortality ratios for each unit. The use of logistic regression models provided a powerful and flexible methodology that allowed risk-adjustment.

Standardized mortality ratio and confidence interval

The results were expressed as standardized mortality ratios (SMRs). The SMR has an intuitive interpretation that is widely understood. The two main difficulties with this approach are that SMRs cannot be directly compared between units, and the lack of an accepted method to calculate confidence intervals. Section 5.5.3 explored the first problem, noting that, in practice, any bias introduced by population differences among the units is usually small. Methods to estimate confidence intervals for the SMR were explored through a simulation study and by applying them to the TNS data. The method used subsequently in this thesis (Hosmer and Lemeshow, 1995) was shown to have adequate coverage properties.

Parameterization of the reference units

Deviation contrasts were used for the reference units to allow the estimation of indirectly standardized mortality rates for each unit. This parameterization reduced the influence of the larger units, while ensuring the most efficient estimation of covariate effects.

7.2.3 Main results

Although the primary aim of the thesis was to identify appropriate statistical methods, risk-adjusted mortality rates were estimated for the units. The main results are set out in this Section.

Frequentist analysis

The two main models from this thesis are the 'full' and 'reduced' models from Chapter 6. In these models, neither of the estimated 95% confidence intervals for the SMR associated with Unit 6 contained the value one. No other unit had a 95% confidence interval that did not contain unity for either the 'full' or the 'reduced' model. These models provided evidence that Unit 6 has a risk-adjusted mortality rate higher than the other units. This conclusion held true both when an alternative parameterization of the reference units was specified and when Apgar score at one minute was removed from the model.

Bayesian analysis

The values of the estimated SMRs, and 95% confidence intervals, were similar between the 'reduced' models using classical and Bayesian approaches, although in the Bayesian analysis the 95% credible interval for Unit 14 did not contain the value one. However, probability statements about the SMRs were available from the Bayesian models. The probability that the true value of the SMR was greater than $\frac{3}{2}$ was over 0.5 for Units 4. In other words, it was more likely than not that their true SMR was over $\frac{3}{2}$. There were five units where the probability that the true SMR was under $\frac{2}{3}$ was more than 0.5: Units 8, 9, 12, 13 and 14.

Results for Unit 9

Unit 9 was not included in the main frequentist models of Chapter 6: 'full' and 'reduced'. The only infant admitted to Unit 9 who died before admission had no observed value for Apgar score at one minute, one of the variables in the final risk-adjustment model. Since all of the observations from that unit that could be included in the model then had the same outcome, quasi-complete separation of the data occurred and MLE for the model parameters became unstable (§3.4). The exclusion of Unit 9 from the main analyses may have meant it not being identified as having an extreme mortality rate. Although it is difficult to imagine a scenario where a unit with only one recorded death is identified as having a high mortality rate, a low number of deaths may indicate a low rate. However, a model without Apgar score at one minute, but including Unit 9, was reported in §6.9. The estimated SMR showed a wide confidence interval for Unit 9 and no evidence for a particularly low SMR.

7.3 Discussion

7.3.1 Clinical importance of the thesis

The desire to monitor the performance of health care providers is understandable. This thesis has illustrated, discussed and developed methods to compare the rates of in-unit mortality in sixteen neonatal intensive care units.

Although the discussion and application of the statistical methods has emphasised their use in neonatal medicine, these methods are applicable to all medical disciplines and to a wide range of binary performance indicators, be they measures of process, structure or outcome. Much of the thesis is also relevant to indicators measured on other scales, for example categorical,

ordinal or continuous. The need for adequate risk-adjustment holds for all provider profiling and results need to be reported in a clinically useful manner that can be understood by all potential users.

7.3.2 Comparison with previous studies

The aim of the thesis was to identify the most appropriate statistical methods to allow the reporting of in-unit mortality rates of neonatal units compared to the other units in the region. Provider profiling is undertaken by many organizations in a range of medical specialities. In this Section the analyses illustrated in this thesis are compared to recent high-profile profiling exercises and to the current reporting of mortality rates in the Trent Neonatal Survey annual report.

Trent Neonatal Survey annual report

Currently, the analyses for the TNS annual report use indirect standardization through logistic regression models and report the SMR for each unit. Until 2002 the observed and expected number of deaths were reported, but since then the standardized mortality ratio has been published. The models use the 'rest of Region' parameterization described in §5.3.1 and 95% confidence intervals are estimated using the Normal approximation method, without the continuity correction (The Trent Infant Mortality and Morbidity Studies, 2003).

The risk-adjustment method is the CRIB II score (§4.4.1), without the component for temperature at admission, as discussed above. The TNS annual report contains four analyses of in-unit mortality by neonatal unit, each undertaken using a different population: i) 20-32 weeks gestational age; ii) 20-32 weeks gestational age excluding post-natal transferred infants; iii) 25-32 weeks gestational age; iv) 25-32 weeks gestational age excluding post-natal transferred infants. Such sub-group analyses have not been undertaken in this thesis, but the methods illustrated here can be used for such investigations.

Bristol Royal Infirmary inquiry

From 1998 to 2001 a public inquiry was held to investigate mortality following paediatric cardiac surgery at Bristol Royal Infirmary (Department of Health, 2000). Concerns had previously been raised of a high death post-operative rate (Delamothe, 1998). The inquiry included statistical analysis of mortality rates at Bristol compared to similar centres in the UK. The statistical analysis of these data comprised three analyses of increasing complexity: referred to as 'one'-, 'two'- and 'three'-star analyses (Spiegelhalter et al. 2002). The

'one'-star analysis was equivalent to the 'rest of region' parameterisation described in §5.3.1, in that outcomes at Bristol Royal Infirmary were compared to the pooled outcomes from the other centres. Their 'two'-star analysis was a fixed effects logistic regression model with a term for each centre. Such a model was not illustrated in this thesis, as the weighted models illustrated were felt to be superior as they reduced the influence of the largest units. The 'three'-star model developed for the analysis was a hierarchical logistic model that allowed heterogeneous between-centre variability across different risk-adjustment strata. Hierarchical modelling was discussed in §5.10.

These analyses used routine hospital episode statistics, which only allowed risk adjustment by the type of operation performed. Such analyses have drawn criticism (Gibbs et al. 2002), but there is some evidence that this form of adjustment is sufficient for cardiac surgery (Jenkins et al. 2002; Aylin et al. 2005). Aylin *et al* had previously found that hospital episode statistics had given similar results to analyses using the UK cardiac surgical register (Aylin *et al*, 2001a). However, such statistics are likely to be insufficiently detailed for neonatal intensive care, making a data collection system such as TNS necessary.

Although Bayesian methodology was employed, little attention was given to the usefulness of Bayesian posterior probabilities. As has been demonstrated in this thesis, their use can encourage the use of clinical, rather than statistical, criteria to judge providers.

Adult Cardiac Surgical

The largest on-going profiling exercise is that undertaken by the New York State Department of Health (New York State Department of Health, 2004), which started collecting and publishing information on mortality rates after coronary artery bypass surgery in 1989. Rates are published for both hospitals and individual surgeons. Data are collected for the profiling and risk-adjustment is through a study-specific model. The results are reported as **risk-adjusted mortality rates** (RAMR), which are SMRs multiplied by the overall statewide mortality rate. Ninety-five percent confidence intervals are also reported, although it is unclear how these are estimated. The methodology used closely corresponds to that applied in this thesis. The major difference is that the SMR is multiplied by the overall mortality. This, it is claimed, *"...is the best estimate, based on the statistical model, of what the provider's mortality rate would have been if the provider had a mix of patients identical to the statewide mix."* (New York State Department of Health, 1998a). However, it has been shown in this thesis that this is not necessarily true as a statistic based on indirect standardization,

such as the SMR and RAMR, are weighted by the providers case-mix. The reporting of the RAMR only serves to veil this difficulty.

As in the USA, adult cardiac surgery has taken the lead in the publication of the performance of individuals (Society of Cardiothoracic Surgeons of Great Britain and Ireland, 2005). From 2005, 30-day mortality rates of individual surgeons following cardiac surgery will be made public. These will be published annually for heart surgery centres and for individual surgeons on a rolling three-year basis (Department of Health, 2002a)117).

While the methods to be used for the national reporting of results are currently unknown, several hospitals have independently published risk-adjusted mortality rates for surgeons. However, the presentation of these results has not been consistent. The Manchester Heart Centre (Manchester Heart Centre, 2005) presented the data as cusum plots, with the Parsonnet score used for risk adjustment (Parsonnet *et al,* 1989). St George's Hospital, London presented the results for individual surgeons as cumulative plots of the difference between the observed and expected mortality, without confidence intervals (St George's Hospital, 2005). In this case, the EuroSCORE was used for risk-adjustment (Roques *et al,* 2003). The results for Papworth Hospital, Cambridge also used the EuroSCORE but compared rates of observed and expected death for an entire financial year (Papworth Hospital, 2005).

Although the approach used by Papworth Hospital most closely matches that presented in this thesis, the reporting of indicators for individual surgeons differs from that of TNS. First, the results for surgeons are more likely to be available continuously, allowing the used of methods such as cusum plots and VLAD (§3.5). Although the reporting of results for individual surgeons is not uncontroversial, since other factors affect the outcome of a patient, i.e. the anaesthetist, intensive care medical and nursing staff (Keogh *et al,* 2004), the surgeon has a level of impact on the patient that a neonatal consultant does not (Field *et al,* 2002). Therefore, although some of the issues discussed in this thesis are relevant to the profiling of cardiac surgeons (e.g. risk-adjustment, confidence intervals, presentation), the most appropriate statistical method may be different.

Dr Foster

Other work that receives a large amount of attention is the information published annually by the Dr Foster organization (Dr Foster, 2004) and reproduced in the Sunday Times newspaper (Sunday Times, 2004b). Their analyses are based on aggregated routine data rather that purposely collected information. The use of data collected for purposes other than profiling

can be unreliable (Bridgewater *et al,* 2002). In addition, aggregated data are used to construct linear regression models to estimate SMRs, called a *"Mortality Index"* (§3.5). However, it has long been known that correlations at a group level may not represent correlations at the individual level: the ecological fallacy (Robinson, 1950; Selvin, 1958). Therefore, since TNS data are collected on individuals, the methods in this thesis are superior to that employed by the Dr Foster organisation.

7.3.3 Implication for policy

Usefulness of provider profiling

The ultimate question is whether health care provider profiles, such as that outlined in this thesis, are useful: for example to staff, patients or funders. The general answer must be that profiling has an important role in allowing providers, and others, to see where their performance (or at the very least, outcomes) lies in relation to a reference. Without this, the quality of the care provided cannot be adequately assessed. However, while profiling is desirable in principle, it is not as clear whether this is true in practice. Such profiles have many potential difficulties and it must be determined whether these problems result in any conclusions being meaningless, or worse, misleading.

To quantify the quality of care offered in a NICU, the measure to be investigated should be both specific and sensitive to the care given. Although mortality is likely to be reliably recorded, it is unclear whether it is the most desirable outcome in all cases. However, the methods considered in this thesis are applicable to any choice of binary indicator.

The effectiveness of the retrospective profiling of neonatal intensive care units has been questioned (Parry *et al,* 1998). Parry *et al* investigated annual mortality rates in nine NICUs in the UK over six years and demonstrated that the variation in rates over time was greater than the variation between units. This, they concluded, meant *"Annual league tables are not reliable indicators of performance or best practice"*. Instead, it was recommended that prospective studies to investigate the association between outcome and units characteristics, such as volume, staffing levels, training and expertise, be considered. While such studies would undoubtedly be useful, the type of profiling set out in this thesis still has a role in monitoring outcomes. Whatever the workload or staffing levels of a unit, the practice and organisation of the neonatal team will nevertheless influence the quality of care that an infant receives. The subtleties of the most appropriate treatment in each circumstance are unlikely to be agreed and some measure of outcome would be necessary (§1.3.1).

Parry *et al* used annual mortality rates, which resulted in small samples for the units: the annual number of infants for all nine units ranged from 389 to 490. The number of infants admitted to individual units in their study ranged from 11 to 127 infants, with between 0 and 25 observed deaths. Such small numbers are likely to display large random variation and three-year averages were used in this thesis to overcome the problem. However, the use of data covering several years may mean that results are out-of-date when published and unlikely to reflect current practice. Whether this is true depends on any changes in practice or organisation over that time. In either case, if such profiling exercises are seen as opportunities to identify best practice then the identification of the characteristics of the best performers can aid the improvement of all units. If substantial changes had occurred in units then the effect of these changes on clinical outcomes can be assessed.

Identification of units with extreme rates

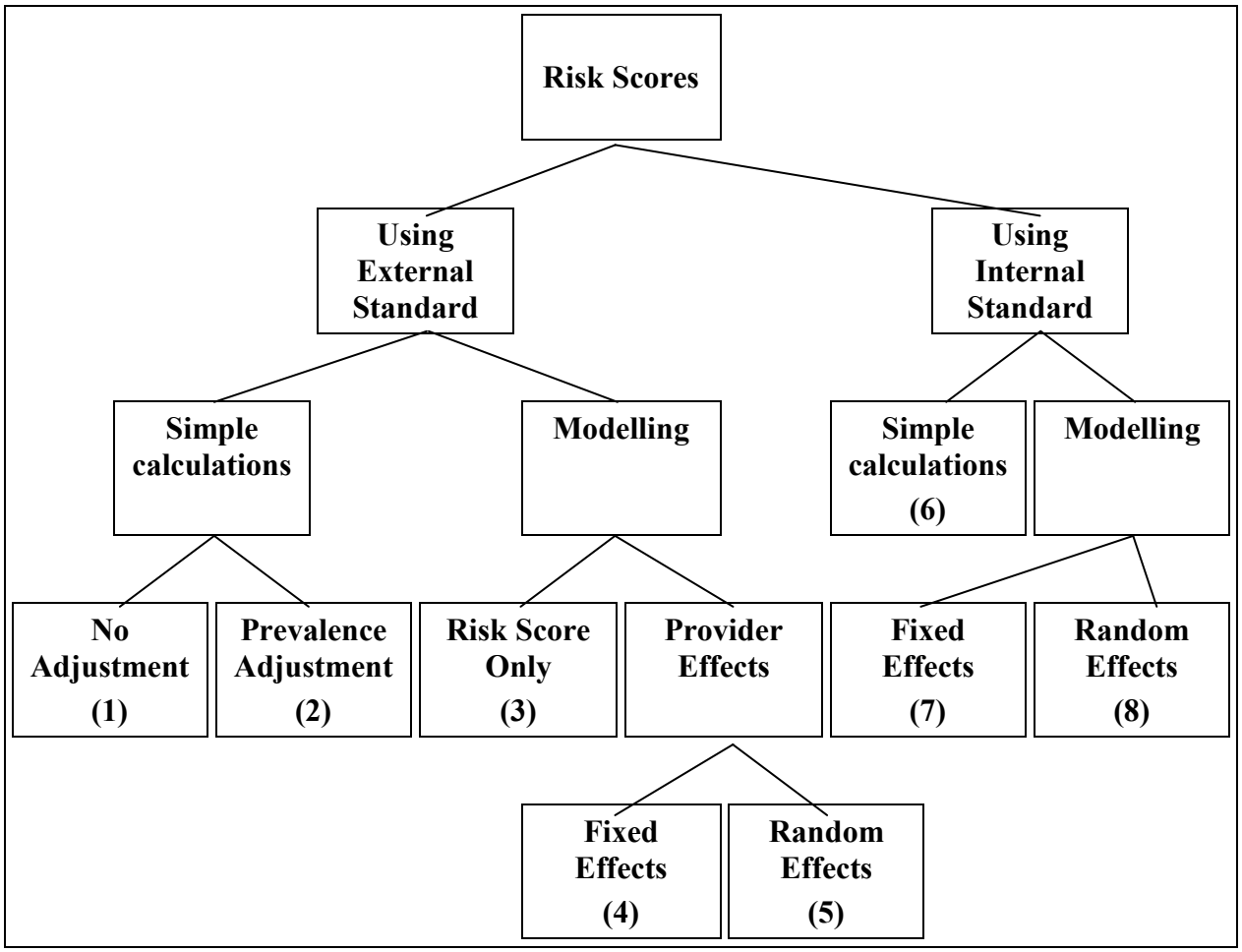
In this thesis Unit 6 has been identified as potentially having a high risk-adjusted mortality rate. This unit had been identified from previous TNS annual reports and an internal investigation into clinical practices is underway. In addition, the Bayesian analysis in this thesis has identified other units, including three with potentially low mortality rates that may be worthy of investigation. However, improvements to the reduced model are required to confirm these results.

7.3.4 Statistical considerations

The methods outlined in this thesis are all based on statistical analyses. The relevant data are collected, statistical methods applied and then inferences made about the neonatal units. However, this is not the only approach that could have been taken. It could be argued *"... that a definitive assessment of quality must be based on a knowledge of all the particulars in a case, so that an assessor recognised to have superior skill can reconstruct in his own mind the conduct of care that he would have recommended under the circumstances."* (Donabedian, 1978) While it may be possible that such an approach using a review of notes may be possible for a relatively small number of cases (such as is undertaken by CEMACH into maternal deaths), to carry out such a task for in-unit deaths for an entire Region is probably impossible in terms of time and cost, let alone the agreed definition of the appropriate method of care in all circumstances. Thus, a statistical approach is likely to be the only one that can practically be taken.

DeLong *et al* (1997) identified eight statistical methods that may be useful in provider profiling. These methods were identified and numbered in a flow-chart: reproduced here as Figure 7.1.

Figure 7.1 Possible risk-adjustment methodologies



Methods 1 to 5 use an externally derived risk-adjustment algorithm, either to directly obtain the expected probability of death for each infant (methods 1 & 2) or to use the value of the external linear predictor as a covariate in a new logistic regression model (methods 3, 4 & 5). No suitable risk-adjustment scores existed for the TNS data (discussed in Chapter 4) so these approaches were not pursued in this thesis. However, methods 6, 7 and 8 would have been possible with the TNS data as they involve using internal data to risk-adjust. DeLong *et al*'s method 6 is analogous to the 'rest of Region' parameterisation discussed in §5.3.1. For method 7 indicator variables for the providers are included in the model together with the risk-adjustment variables. The final model, model 8, is a random effects approach pursued in §5.10 and discussed further below.

Random effects models

Although illustrated in §5.10, it was felt that the reduced true positive rate encountered with such models was inappropriate for the profiling exercise in this thesis. The aim of these analyses was to identify units that may potentially have extreme mortality rates and may benefit from further, clinical based, investigations. The units were anonymised, as they are in the Trent Neonatal Survey annual reports, to eliminate the fear that a unit would be identified as an outlier before a full clinical investigation could be undertaken.

Propensity score

It has been suggested that propensity score methods offer a more robust approach to risk-adjustment in provider profiling (Huang *et al*, 2005). Propensity score methods were developed in econometrics, where they are often known as the 'Hickman Adjustment' (Hickman and Hotz, 1989), and were introduced into medical research by Rubin (1997). The method comprises the estimation of predicted probabilities of being in the 'intervention' group using the risk-adjustment variables: the propensity score. The observations are then divided into groups (usually quintiles) and the observations compared both within these groups and by the estimation of an overall, weighted, effect. Such methods are particularly useful where there is little overlap in the values of covariates between the two 'intervention' groups.

While this method is potentially very useful in provider profiling, it was not considered useful for the data in this thesis. In particular, the small number of observations in some units meant that there would be insufficient data to apply the methods.

Residual confounding

As with all observational studies, there is the potential for confounding to remain in the final models, and to be the reason for extreme risk-adjusted mortality rates observed. The morbidity of a newborn is a highly complex phenomenon that statistical models are very unlikely to be able to capture exactly. However, the statistical model used does not have to be accurate in that it replicates the biological processes within a mathematical framework, although it does help if it is clinically plausible. Rather, the statistical model should adequately predict the outcomes modelled. The final models used in this thesis demonstrated high discriminatory ability ('full' model $AROC$ = 0.94; 'reduced' model $AROC$ = 0.92) and good calibration, both as measured by the Hosmer & Lemeshow goodness-of-fit test ('full' model p = 0.17; 'reduced' model p = 0.17) and by calibration plots.

Multiple Testing

The issue of multiple significance testing has not been raised, so far, in this thesis. The aim of the analyses in this thesis is not to compare each unit with every other, giving therefore ${}_{16}C_2 = 120$ comparisons. Rather, the 16 comparisons between each unit and the others as a whole were of interest. It can be argued that in such circumstances the true experiment-wise Type I error rate was $1-(1-0.05)^{16} = 0.56$ (Motulsky, 1995:120). The simplest approach to adjusting the Type I error rate is to use a Bonferroni correction and divide the required experiment wide *P*-value by the number of comparisons. Alternatively, to ensure a true experiment-wise Type I error rate of 0.05, each comparison-wise significance probability should have been $1-\sqrt[16]{(1-0.05)} = 0.0032$ and 99.68% confidence intervals estimated. However, such an approach means that the statistical significance of a comparison will depend on the number of other units being considered (Perneger, 1998). In this thesis, the global hypothesis that none of the units differ from the rest of the Region is not the hypothesis of interest and the 5% significance level should not be applied to this hypothesis (DeLong et al. 1997). In addition, any reduction in the rate of Type I errors results in an increase in the rate of Type II errors.

7.4 Limitations of the Study

Confidence interval estimation for SMR

In the thesis the standardized mortality ratio (SMR) was proposed as the most appropriate summary statistic. However, the lack of an accepted method to obtain a confidence interval for the SMR was described in §5.6 and the properties of various methods were explored. The method felt to offer the best coverage properties, the bootstrap method, was found to be inappropriate when there were small numbers of outcomes, and the method proposed by Hosmer and Lemeshow was used instead. Although this method demonstrated good coverage rates (close to the nominal 5% level) the intervals were often one-sided with high values for the upper limits, especially for small units. This may lead to small units with low mortality rates not being identified.

Measure of the quality of care?

It is important that any profiling answers the correct question and different users will have different questions. Perhaps at its most basic, clinicians may be most interested in processes and patients interested in outcomes. In other works, clinicians want to know that they are doing the right thing and patients want to know that they are getting the best outcomes. It is arguable that measuring mortality does not provide an answer to either of these questions. Death rates may not sufficiently reflect the quality of the care given, and survival at any costs may not be the best outcome in all circumstances. It is vital not to mimic the drunk in the joke who searches for his lost keys under a street lamp rather than where he lost them because, he rationalises, "There is more light over here." Mortality may be an easy indicator to monitor but, in neonatal medicine, survival at all costs may not be a sensible goal.

Missing values

Although the Trent Neonatal Survey ensures high quality data collection through the employment of specialist neonatal nurses, some values were missing from the data. In particular, values of maximum base excess in the first 12 hours of life and Apgar score at one minute of life were missing from a large proportion of observations (24.7% and 4.6% respectively). Observations with missing values for base excess were assumed to have values within the normal range. It was felt that in most cases the value had not been calculated as it was felt to be unremarkable. However, this assumption has not been validated and departure from this assumption may bias the results.

Observations with missing Apgar scores were excluded from analyses when necessary. The reasons that the Apgar scores were missing are unknown, as a score is routinely given to each newborn infant. The major effect on excluding these observations was the need to exclude Unit 9 from such analysis as the only observed death had a missing Apgar score. The introduction of any biases into the model through the exclusion of these observations is unknown (see §7.5).

Poor care or deliberate actions

The methods described in this thesis are proposed to try to identify 'good' and 'poor' clinical practice. As such, they are not designed to try to identify acts of murder by health care workers. Unfortunately, the existence of health care professionals who murder is not unknown: perhaps, as has been suggested, health care produces more serial killers that all

other professions put together (Kinnell, 2000). Two recent cases have received a lot of media attention.

In 2000 the GP Harold Shipman was convicted of the murder of 15 of his patients although the official enquiry concluded that Shipman had murdered 215 patients, with a *"real cause to suspect"* that he might have been responsible for a further 45 deaths (The Shipman Enquiry, 2002:197). These murders were carried out from March 1975 to June 1998.

The second recent case is that of Beverly Allitt, a nurse convicted in 1993 of murdering four children, attempted murder of a further three and causing grievous bodily harm to another six. All of these attacks took place on the children's ward of Grantham and Kesteven Hospital in Lincolnshire over 58 days from February to April 1991 (Dyer, 1994).

These two cases are very different: one a GP for the most part in a single-handed practice who killed for over 20 years, the other in a hospital for less than two months. The procedures required to identify such behaviour in the future are likely to depend on the setting (Baker et al. 2003; James and Leadbeatter, 1997). Although the methods set out in this thesis may pick up such murders they are unlikely to be either the most effective or efficient ways.

7.5 Further Work

Analyses of sub-groups

This thesis has used a single summary statistic (the SMR) to describe the deviation of a unit's risk-adjusted mortality from the rest of the units in the Region. However, the use of a single statistic may hide extreme rates in sub-groups of infants: for example, a high rate in one group could be balanced by a low rate in another leading to an unremarkable SMR. The analyses of clinically important sub-groups can help to identify this type of scenario. For units with extreme values for the overall SMR, such sub-group analyses can also help to identify for which infants their outcome differs from the other units.

Random effects modelling

The debate over the choice of fixed or random effects modelling is unresolved. The results from any TNS analysis of mortality rates are anonymous to all but the relevant neonatal unit, as these results are intended as indicators for further investigation. In these circumstances, the likely reduction in the false positive rate, observed from a random effects model, would not

compensate for the missing opportunities that result from the increased false negative rate (§5.10).

The choice of model should take into account the costs associated with each type of error (Draper and Gittoes, 2004) and further work is required as the relative sizes of these error rates are unknown for these data. Currently, the NICUs are anonymised when mortality rates are reported (The Trent Infant Mortality and Morbidity Studies, 2003), as in this thesis. However, it is unknown whether such a policy will be permitted if named data are requested under the Freedom of Information Act 2000. If named data are to be published there may be pressure to use statistical methodology that reduces the Type I error rate, such as random effects modelling.

Bayesian models

The Bayesian model described and illustrated in Chapter 6 produced results similar to those from the frequentist analysis but also allowed probability statements to be made about the SMR. A further advantage of Bayesian methods is the ability to formally include prior information into the model. Further investigate into the choice of prior distributions for the model parameters and their influence on the parameter estimates would shed light on the usefulness of such models in provider profiling.

Variable selection for a reduced model

Although the reduced model developed in Chapter 6 showed good calibration and discrimination, it was suggested in §7.3.4 that important variables might be missing. When compared to the results from the 'full' model, there was an excess of small units with low estimated values for the SMR. Further work is required to find which variables, when included in the reduced model, can overcome this problem.

If such a model can be obtained it may be useful to develop it into a general mortality risk-adjustment score. Problems with existing scores were discussed in Chapter 4. To be of use, such a score must be appropriate for populations other than TNS. It is unlikely that a data-driven model selection method, such as forward selection by statistical significance as used in this thesis, would produce such a generalizable model (Harrell, 2001:56-60). A more sophisticated approach would be more likely to produce a reliable model (Harrell, 2001:79-82). Such a model could then be converted to a clinical score (Cole, 1993). However, data other than from TNS would be required to investigate the generalizability of the model.

Zero observed deaths

The difficulties encountered when a provider presents with no observed events were discussed in §6.11. The statistical methodology used here resulted in Unit 9 being excluded from the final frequentist analyses (§6.5 & 6.6), both as part of the reference population and as the unit investigated. Although various solutions were suggested for both situations, each of these has limitations. Further work is required to identify the most appropriate methods (both Bayesian and Frequentist) for the analysis of such data, as such units are likely to occur more frequently with the wider adoption of profiling.

7.6 Conclusions

The aim of this thesis was to identify, critically appraise, and where appropriate develop, statistical methods appropriate for the analysis of in-unit mortality for the TNS data. The use of standardized mortality ratios as the summary statistic is entirely appropriate. Although direct comparisons between the units are not the primary aim of the analysis, there is no evidence that such comparisons would result in grossly misleading conclusions. The full risk-adjustment model showed both excellent discrimination and calibration.

A Bayesian approach has been demonstrated that offers advantages of being able to both formally include prior information and to derive probability statements for the SMRs. However, the sensitivity of the estimates to changes in the prior distributions needs to be investigated further before this approach is considered for the TNS annual report.

Although the identification of outlying units was not the primary aim of this thesis, evidence has been show that Unit 6 had a high death rate that was statistically significant at the 5% level. The Bayesian model also showed that Units 8, 9, 12, 13 and 14 (low) and Unit 4 (high) might also merit further investigation.

Recommendations for future TNS annual reports

The current method of reporting of death rates in the TNS annual reports was described in §7.3.2. This thesis has shown that the reporting of mortality through SMRs is appropriate. However, the use of deviation parameterizations of the reference units would remove the greater influence of the larger units and ensure that each unit has equal influence on the reference mortality rates. In addition, it has been shown that the use of the Normal approximation method to estimate confidence intervals is likely to mean previously reported

intervals are either too large or too small for those units with SMRs away from unity. The use of the method proposed by Hosmer & Lemeshow would produce more reliable confidence intervals. These two changes are simple to implement for the annual report and are recommended to be implemented in the next report.

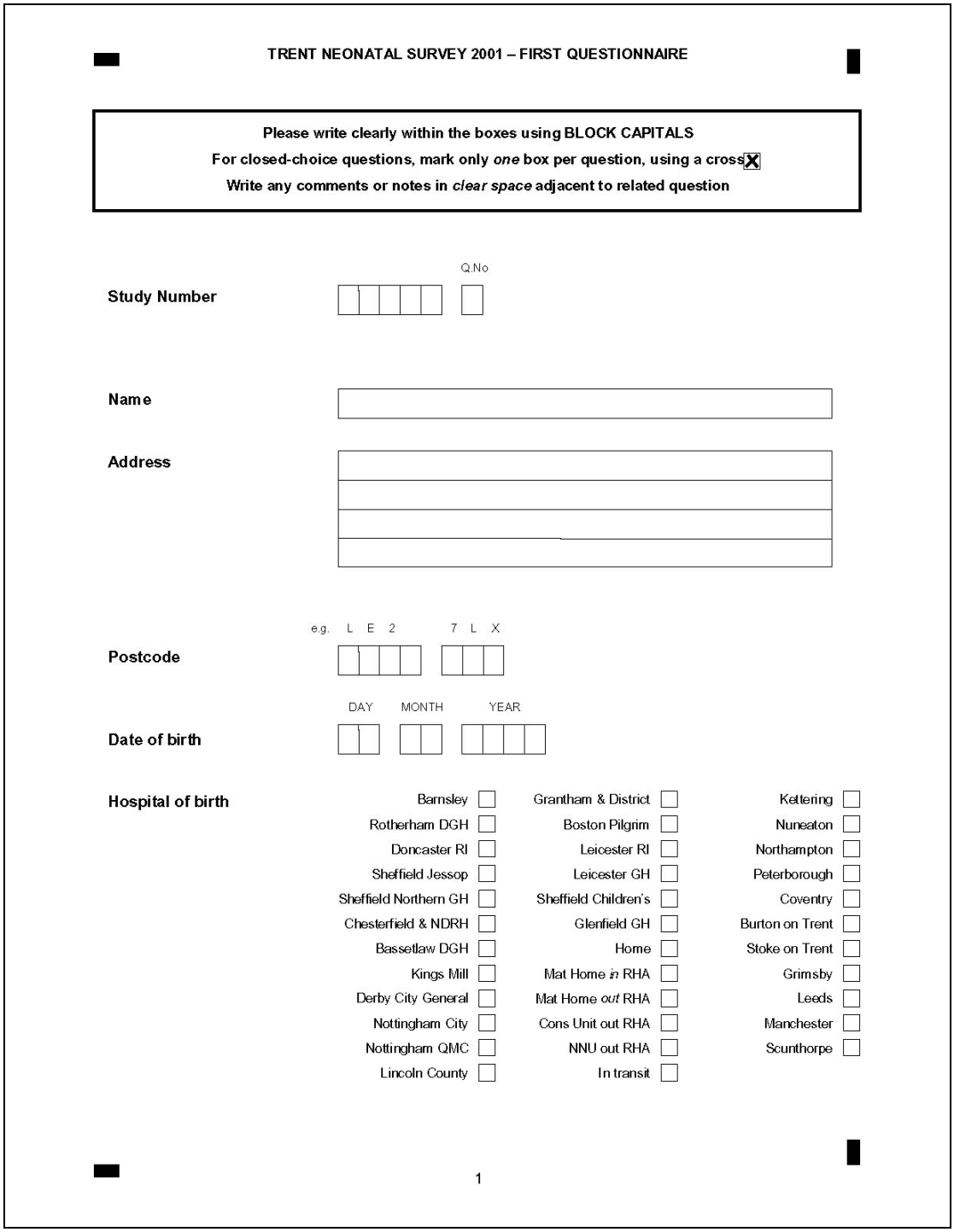
The risk-adjustment proposed in the 'reduced' model is similar to the method currently utilized for the TNS annual report. The current method is likely to be sufficient until further work has been undertaken on the 'reduced' model.

The use of Bayesian methods has been shown to advantageous, particularly as probability statements can be made concerning the estimated standardized mortality ratios. Further work is required into the choice of prior distributions and a recommendation that such methods be used in the TNS annual reports is withheld until more work has been carried out. However, should these methods prove to be robust, a Bayesian approach should be applied to the analysis of TNS data.

APPENDICES

Appendix A: TRENT NEONATAL SURVEY QUESTIONNAIRE

Figure A.1 Trent Neonatal Survey Report Form



TRENT NEONATAL SURVEY 2001 – FIRST QUESTIONNAIRE

Please write clearly within the boxes using BLOCK CAPITALS
For closed-choice questions, mark only *one* box per question, using a cross X
Write any comments or notes in *clear space* adjacent to related question

Q.No

Study Number

Name

Address

e.g. L E 2 7 L X

Postcode

DAY MONTH YEAR

Date of birth

Hospital of birth

Barnsley	Grantham & District	Kettering
Rotherham DGH	Boston Pilgrim	Nuneaton
Doncaster RI	Leicester RI	Northampton
Sheffield Jessop	Leicester GH	Peterborough
Sheffield Northern GH	Sheffield Children's	Coventry
Chesterfield & NDRH	Glenfield GH	Burton on Trent
Bassetlaw DGH	Home	Stoke on Trent
Kings Mill	Mat Home *in* RHA	Grimsby
Derby City General	Mat Home *out* RHA	Leeds
Nottingham City	Cons Unit out RHA	Manchester
Nottingham QMC	NNU out RHA	Scunthorpe
Lincoln County	In transit	

1

TRENT NEONATAL SURVEY 2001 – FIRST QUESTIONNAIRE

Hospital of this admission

Barnsley ☐	Grantham & District ☐	Kettering ☐
Rotherham DGH ☐	Boston Pilgrim ☐	Nuneaton ☐
Doncaster RI ☐	Leicester RI ☐	Northampton ☐
Sheffield Jessop ☐	Leicester GH ☐	Peterborough ☐
Sheffield Northern GH ☐	Sheffield Children's ☐	Coventry ☐
Chesterfield & NDRH ☐	Glenfield GH ☐	Burton on Trent ☐
Bassetlaw DGH ☐	Home ☐	Stoke on Trent ☐
Kings Mill ☐	Mat Home *in* RHA ☐	Grimsby ☐
Derby City General ☐	Mat Home *out* RHA ☐	Leeds ☐
Nottingham City ☐	Cons Unit out RHA ☐	Manchester ☐
Nottingham QMC ☐	NNU out RHA ☐	Scunthorpe ☐
Lincoln County ☐	In transit ☐	

Date of admission DAY ☐☐ MONTH ☐☐ YEAR ☐☐☐☐

Method of delivery

Normal ☐	Low forceps ☐	High forceps ☐
Ventouse ☐	Assisted breech ☐	Emergency CS - labouring ☐
Emergency CS - not labouring ☐	Elective CS ☐	N/K ☐

Birth weight ☐☐☐☐ grams

Gestation ☐☐ weeks

Reason for admission

Respiratory ☐	Pre-term infant ☐	IUGR ☐
Jaundice ☐	Surgery ☐	Con. anom. or inherited disorder ☐
Infection ☐	Observation ☐	Severe neurological disturbance II ☐
Severe neurological disturbance III ☐		

Congenital anomaly

Nil ☐	Chromosomal ☐	Genito-urinary ☐
Neural tube defect ☐	Other neurological lesions ☐	GIT upper (above diaphragm) ☐
GIT lower (below diaphragm) ☐	Craniofacial ☐	Cardiac ☐
Complex / multiple ☐	Other ☐	

Is the condition lethal? Yes ☐ No ☐ N/A ☐ N/K ☐

Days in oxygen ☐☐☐

Days in CPAP ☐☐☐

Days on ventilator ☐☐☐

2

TRENT NEONATAL SURVEY 2001 – FIRST QUESTIONNAIRE

Days of TPN ☐☐☐

Date of discharge / death DAY ☐☐ MONTH ☐☐ YEAR ☐☐☐☐

Discharged to

Barnsley ☐	Grantham & District ☐	Kettering ☐
Rotherham DGH ☐	Boston Pilgrim ☐	Nuneaton ☐
Doncaster RI ☐	Leicester RI ☐	Northampton ☐
Sheffield Jessop ☐	Leicester GH ☐	Peterborough ☐
Sheffield Northern GH ☐	Sheffield Children's ☐	Coventry ☐
Chesterfield & NDRH ☐	Glenfield GH ☐	Burton on Trent ☐
Bassetlaw DGH ☐	Home ☐	Stoke on Trent ☐
Kings Mill ☐	Mat Home *in* RHA ☐	Grimsby ☐
Derby City General ☐	Mat Home *out* RHA ☐	Leeds ☐
Nottingham City ☐	Cons Unit out RHA ☐	Manchester ☐
Nottingham QMC ☐	NNU out RHA ☐	Scunthorpe ☐
Lincoln County ☐	In transit ☐	Birmingham Children's ☐
General paediatric ward ☐	Post-natal ward ☐	Death ☐

Length of stay ☐☐☐ days

Type of transfer None ☐ In utero ☐ Flying squad ☐ Routine ☐ N/K ☐

Sex Male ☐ Female ☐ N/K ☐

Indicate multiplicity of birth 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6+ ☐ N/K ☐

Surfactant? Yes ☐ No ☐ N/K ☐

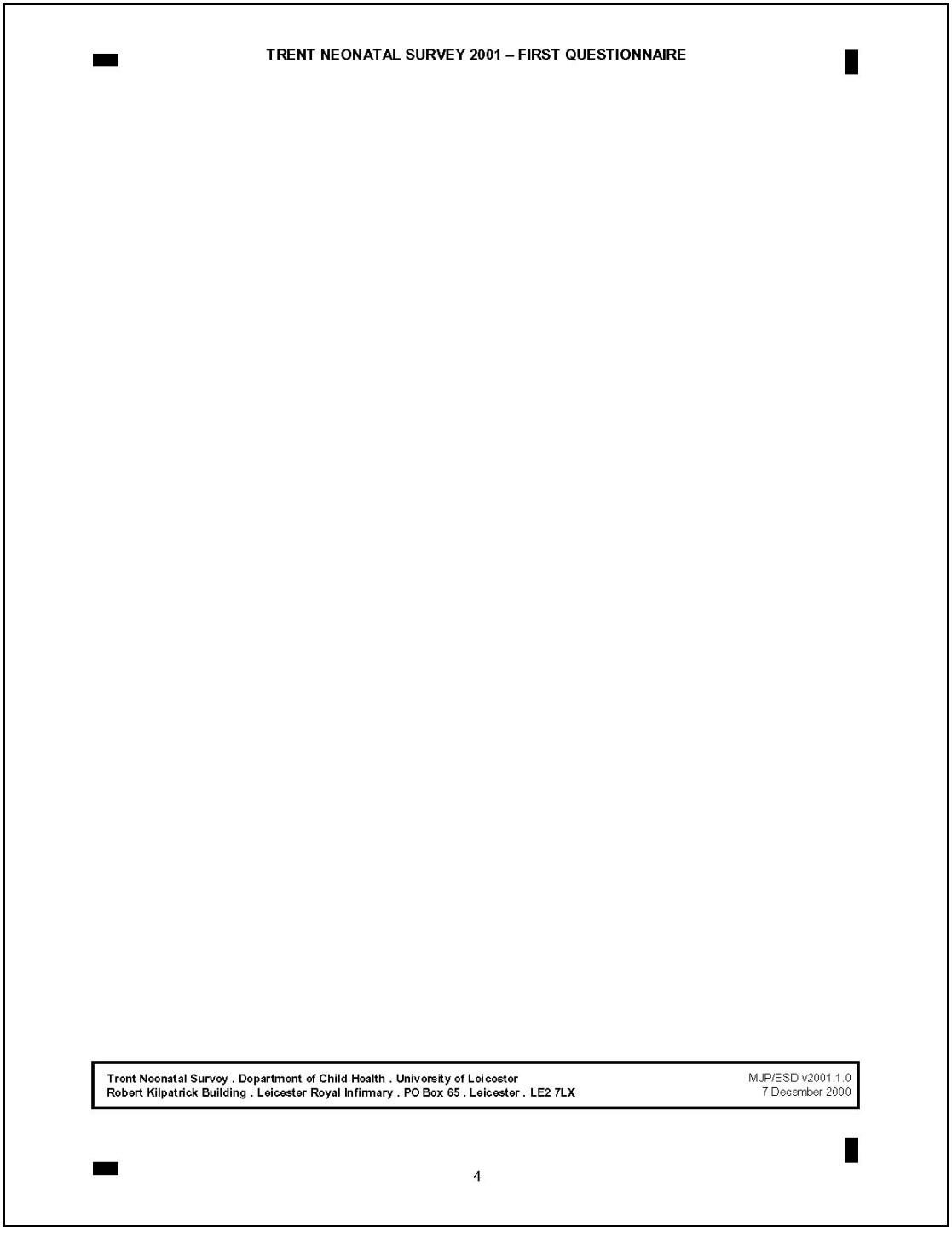
Antenatal steroids at any time prior to labour? Yes ☐ No ☐ N/K ☐

HIGHEST FiO_2 in first 12 hours after resuscitation ☐☐☐

LOWEST FiO_2 in first 12 hours after resuscitation ☐☐☐

Maximum base excess in first 12 hours
(enter 999.9 if not measured)
± ☐☐☐☐.☐

3



TRENT NEONATAL SURVEY 2001 – FIRST QUESTIONNAIRE

Trent Neonatal Survey . Department of Child Health . University of Leicester
Robert Kilpatrick Building . Leicester Royal Infirmary . PO Box 65 . Leicester . LE2 7LX

MJP/ESD v2001.1.0
7 December 2000

4

TRENT NEONATAL SURVEY 2001 – FIRST QUESTIONNAIRE

PREVIOUS PREGNANCIES

– **live births** ☐☐

– **stillbirths** ☐☐

– **TOPs** ☐☐

– **other abortions** ☐☐

Maternal age ☐☐ years

Was birth preceded by labour? Yes ☐ No ☐ N/K ☐
(any contractions including niggling, i.e. health professional felt contractions)

If Yes, what was the cervical dilation on admission ☐☐ cm
(enter 99 if not known)

Was labour induced? Yes ☐ No ☐ N/K ☐

Were any of the following conditions present prior to birth?

– **Antepartum haemorrhage** Yes ☐ No ☐ N/K ☐
(excluding show / including spotting)

– **Spontaneous prelabour rupture of membranes** *(definite)* Yes ☐ No ☐ N/K ☐

– **Proteinuric hypertension** Yes ☐ No ☐ N/K ☐
(> +protein and diastolic > 90)

– **Proven fetal / maternal infection** Yes ☐ No ☐ N/K ☐
(treatment given or planned)

– **Other** Yes ☐ No ☐ N/K ☐
(if Yes, please specify)

5

TRENT NEONATAL SURVEY 2001 – FIRST QUESTIONNAIRE

In the week prior to birth, did the mother receive any of the following drugs?

– Corticosteroids *(Dexamethasone)*	Yes ☐	No ☐	N/K ☐
– B-sympathomimetics *(Ritodrine, Salbutimol)*	Yes ☐	No ☐	N/K ☐
– Indomethacin	Yes ☐	No ☐	N/K ☐
– Antibiotics	Yes ☐	No ☐	N/K ☐
– TRH	Yes ☐	No ☐	N/K ☐

LABOUR

	HOURS	MINS
Length of first stage	☐☐☐	☐☐
Length of second stage	☐☐	☐☐
Time between membrane rupture and delivery *(maximum 120 hours)*	☐☐☐	☐☐

INFANT

Apgar score at 1 minute	☐☐
Apgar score at 5 minutes	☐☐

INTRAPARTUM MONITORING

– Fetal distress?	Yes ☐	No ☐	N/K ☐
– CTG abnormality?	Yes ☐	No ☐	N/K ☐
– Doppler abnormality?	Yes ☐	No ☐	N/K ☐
– Abnormal scalp pH?	Yes ☐	No ☐	N/K ☐
– Meconium present?	Yes ☐	No ☐	N/K ☐
– Other *(if Yes, please specify)*	Yes ☐	No ☐	N/K ☐

☐

Ethnic group of infant

European ☐ Asian ☐ African / West Indian ☐
Mixed race ☐ Other ☐ N/K ☐

6

Appendix B: NEONATAL MORTALITY RISK SCORES

Clinical Risk Index for Babies (CRIB) (The International Neonatal Network, 1993)

Factor	Score
Birth weight (g)	
>1350	0
851-1350	1
701-850	4
≤700	7
Gestation (wk)	
>24	0
≤24	1
Congenital Malformation*	
None	0
Not acutely life-threatening	1
Acutely life-threatening	3
Maximum base excess in first 12 h (mmol/L)**	
>-7.0	0
-7.0 to -9.9	1
-10.0 to -14.9	2
≤-15.0	3
Minimum appropriate FiO_2 in first 12 h	
≤0.40	0
0.41-0.60	2
0.61-0.90	3
0.91-1.00	4
Maximum appropriate FiO_2 in first 12 h	
≤0.40	0
0.41-0.80	1
0.81-0.90	3
0.91-1.00	5

* Excluding inevitable lethal malformations.

** For example -3.0 mmol/L scores 0, -16.0 mmol/L scores 3.

Clinical Risk Index for Babies (CRIB) II (Parry et al. 2003b)

Male infants

Birth weight (g)											
2751-3000											0
2501-2750										1	0
2251-2500									3	0	0
2001-2250									2	0	0
1751-2000								3	1	0	0
1501-1750						6	5	3	2	1	0
1251-1500					8	6	5	3	3	2	1
1001-1250		12	10	9	8	7	6	5	4	3	3
751-1000		12	11	10	8	7	7	6	6	6	6
501-750	14	13	12	11	10	9	8	8	8	8	
251-500	15	14	13	12	11	10	10				
	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**

Female infants

Birth weight (g)											
2751-3000											0
2501-2750										1	0
2251-2500									2	0	0
2001-2250									1	0	0
1751-2000								3	1	0	0
1501-1750						6	4	3	1	0	0
1251-1500					7	5	4	3	2	1	1
1001-1250		11	10	8	7	6	5	4	3	3	3
751-1000		11	10	9	8	7	6	5	5	5	5
501-750	13	12	11	10	9	8	8	7	7	7	
251-500	14	13	12	11	10	10	10				
	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**

Factor	Score
Temperature at admission (°C)	
≤29.6	5
29.7 to 31.2	4
31.3 to 32.8	3
32.9 to 34.4	2
34.5 to 36.0	1
36.1 to 37.5	0
37.6 to 39.1	1
39.2 to 40.7	2
≥40.8	3
Maximum base excess in first hour (mmol/L)**	
<-26	7
-26 to -23	6
-22 to -18	5
-17 to -13	4
-12 to -8	3
-7 to -3	2
-2 to 2	1
≥3	0

Score for Neonatal Acute Physiology (SNAP) II (Richardson et al, 2001)

Variable	Points	β
SNAP II		
MBP 20-29 mmHg	9	0.88
MBP <20 mmHg	19	1.94
Lowest temperature 95-96ºF	8	0.81
Lowest temperature <95ºF	15	1.55
PO_2/FiO_2 ratio 1.0-2.49	5	0.49
PO_2/FiO_2 ratio 0.3-0.99	16	1.57
PO_2/FiO_2 ratio <0.3	28	2.80
Lowest serum pH 7.10-7.19	7	0.71
Lowest serum pH <7.10	16	1.57
Multiple seizures	19	1.87
Urine output 0.1-0.9 mL/kg/h	5	0.46
Urine output <0.1 mL/kg/h	18	1.82
(Constant)		(-4.69)
Supplemental points to compute SNAPPE II		
Birth weight 750-999g	10	
Birth weight <750g	17	
Small for gestational age (<3rd percentile)	12	
Apgar score at 5 minutes <7	18	

National Therapeutic Intervention Scoring System (NTISS) (Gray et al. 1992)

Item	Subscore
Respiratory	
Supplemental oxygen	1 [a]
Surfactant administration	1
Tracheostomy care	1 [b]
Tracheostomy placement	1 [b]
CPAP administration	2 [a]
Endotracheal intubation	2
Mechanical ventilation	3 [a]
Mechanical ventilation with muscle relaxation	4 [a]
High-frequency ventilation	4 [a]
Extracorporeal membrane oxygenation	4
Cardiovascular	
Indomethacin administration	1
Volume expansion (≤15 mL/kg)	1 [c]
Vasopressor administration (1 agent)	2 [d]
Volume expansion (>15 mL/kg)	3 [c]
Vasopressor administration (>1 agent)	3 [d]
Pacemaker on standby	3 [e]
Pacemaker used	4 [e]
Cardiopulmonary resuscitation	4
Drug therapy	
Antibiotic administration (≤2 agents)	1 [f]
Diuretic administration (enteral)	1 [g]
Steroid administration (postnatal)	1
Anticonvulsant administration	1
Aminophylline administration	1
Other unscheduled medication	1
Antibiotic administration (>2 agents)	2 [f]
Diuretic administration (parenteral)	2 [g]
Treatment of metabolic acidosis	3
Potassium binding resin administration	3
Monitoring	
Frequent vital signs	1
Cardiorespiratory monitoring	1
Phlebotomy (5-10 blood draws)	1 [h]
Thermoregulated environment	1
Noninvasive oxygen monitoring	1
Arterial pressure monitoring	1

Item	Subscore
Central venous pressure monitoring	1
Urinary catheter	1
Quantitative intake and output	1
Excessive phlebotomy (>10 blood draws)	2 [h]
Metabolic/nutrition	
Gavage feeding	1
Intravenous fat emulsion	1
Intravenous amino acid solution	1
Phototherapy	1
Insulin administration	2
Potassium infusion	3
Transfusion	
Intravenous gamma globulin	1
Red blood cell transfusion (≤15 mL/kg)	2 [i]
Partial volume exchange transfusion	2
Red blood cell transfusion (>15 mL/kg)	3 [i]
Platelet transfusion	3
White blood cell transfusion	3
Double blood cell transfusion	3
Procedural	
Transport of patient	2
Single chest tube in place	2 [j]
Minor operation	2 [k]
Multiple chest tubes in place	3 [j]
Thoracentesis	3
Major operation	4 [k]
Pericardiocentesis	4 [l]
Pericardial tube in place	4 [l]
Dialysis	4
Vascular access	
Peripheral intravenous line	1
Arterial line	1
Central venous line	1

Superscript letters represent mutually exclusive variables

National Institute of Child Health and Human Development (NICHHD)

(Hobar et al. 1993a)

$$\textbf{\textit{logit}}(Death) = 2.606 - (0.422 \times BW) - (0.491 \times SGA) - (0.450 \times blackrace) + (0.656 \times male) + (0.971 \times apgar)$$

Where:

BW (birth weight)	per 100g increase
SGA (small for gestational age)	$\begin{cases} 1 & \text{if} < 10^{\text{th}} \text{ centile} \\ 0 & \text{otherwise} \end{cases}$
blackrace (race)	$\begin{cases} 1 & \text{if black} \\ 0 & \text{otherwise} \end{cases}$
male (gender)	$\begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$
apgar (1 minute Apgar score)	$\begin{cases} 1 & \text{if} \leq 3 \\ 0 & \text{otherwise} \end{cases}$

Berlin Score (Maier et al. 1997)

Variable	Range/Grade	Score
Birth weight (g)	1250-1499	3
	1000-1249	6
	750-999	9
	< 750	12
Grade of respiratory distress syndrome	0	0
	I	2
	II	4
	III	6
	IV	8
Apgar score at 5 minutes	> 8	0
	7 to 8	2
	5 to 6	4
	3 to 4	6
	< 3	8
Artificial ventilation	No	0
	Yes	8
Base excess at admission (mmol/l)	≥ -2.0	0
	-2.1 to -5.0	1
	-5.1 to -8.0	2
	-8.1 to -10.0	3
	< -10.0	4

$$\textit{logit}(Death) = -3.7 - (0.84 \times BW) + (0.71 \times RDS) - (0.55 \times APGAR) + (0.53 \times VENT) + (0.37 \times BE)$$

Silkin 12-hour Score (Sinkin et al. 1990)

Variable	**Score (β)**
Birth weight (g)	-1.89
Gestational age (weeks)	-0.21
Apgar score at 5 minutes	-0.25
Peak inspiratory pressure at 12 hours (cm H_2O)	0.13
(Constant)	(8.12)

Neonatal Mortality Prognosis Index (Garcia et al. 2000)

$$\textbf{\textit{logit}}(Death) = -3.1410 + (2.6839 \times [GA \times BW]) + (2.5002 \times CA) + (1.6673 \times [O_2 \times KIRBY]) + (1.0718 \times MCM) + (0.9792 \times S) + (0.8662 \times BE)$$

Where:

GA (gestational age)	$\begin{cases} 1 & \text{if} \leq 32 \text{ weeks} \\ 0 & \text{otherwise} \end{cases}$
BW (birth weight)	$\begin{cases} 1 & \text{if} \leq 1500\text{g} \\ 0 & \text{otherwise} \end{cases}$
CA (cardiac arrest)	$\begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$
O_2 (oxygen saturation)	$\begin{cases} 1 & \text{if} \leq 84\% \\ 0 & \text{otherwise} \end{cases}$
KIRBY (paO_2/FiO_2 ratio)	$\begin{cases} 1 & \text{if} \leq 284 \\ 0 & \text{otherwise} \end{cases}$
MCM (major congenital malformations)	$\begin{cases} 1 & \text{if present} \\ 0 & \text{otherwise} \end{cases}$
S (sepsis)	$\begin{cases} 1 & \text{if present} \\ 0 & \text{otherwise} \end{cases}$
BE (base excess)	$\begin{cases} 1 & \text{if} \leq -10 \\ 0 & \text{otherwise} \end{cases}$

Apgar Score (Apgar, 1953)

	Score		
	0	**1**	**2**
Heart rate	Absent	<100 bpm	>100 bpm
Respiratory effort	Apnoeic	Irregular or shallow breathing	Good
Reflex irritability	Nil	Some	Grimaces, coughing or sneezing
Muscle tone	Flaccid	Good	Spontaneously flexed arms and legs which resist extension
Colour	Blue	Body pink but extremities blue	Pink, including extremities

Transport Risk Index of Physiologic Stability (Lee et al. 2001)

		Score
Temperature:	<36.1 or >37.6	8
	36.1-36.5 or 37.2-37.6	1
	36.6-37.1	0
Respiratory status:	Severe (apnea, gasping, intubated)	14
	Moderate (RR >60/min &/or SpO_2 <85)	5
	None (RR <60/min &/or SpO_2 >85)	0
Systolic BP (mmHg):	<20	26
	20-40	16
	>40	0
Response to noxious stimuli:	None, seizure, muscle relaxant	17
	Lethargic response, no cry	6
	Withdraws vigorously, cries	0

Appendix C: ADDITIONAL DETAILS FROM THESIS

Appendix C contains additional details from the thesis.

First, the choice of Beta(1.25, 3.25) as a prior distribution in §3.4.2 is justified.

Second, Silcock's proof that his *Property 1* does not hold for the standardized mortality rate is reproduced: as discussed on Page 132 (Silcock, 1959).

Next, the derivation of one of the solutions to *Equation 5.31* is given.

Finally, the estimation of the lower 95% confidence limit for Unit 9, using the 'full' normal approximation method, and referred to in §6.11 is shown.

Appendix C.1 Selection of Beta(1.25, 3.25) as a prior distribution for Unit odds

The aim was to find the Beta distribution with the maximum value for the probability $P(\pi<0.02)$, given that the modal value of the distribution was 0.1.

The modal value for a Beta distribution (π_M) is given by:

$$\pi_M = \frac{\gamma - 1}{\gamma + \delta - 2}$$

Hence, in this case:

$$0.1 = \frac{\gamma - 1}{\gamma + \delta - 2}$$

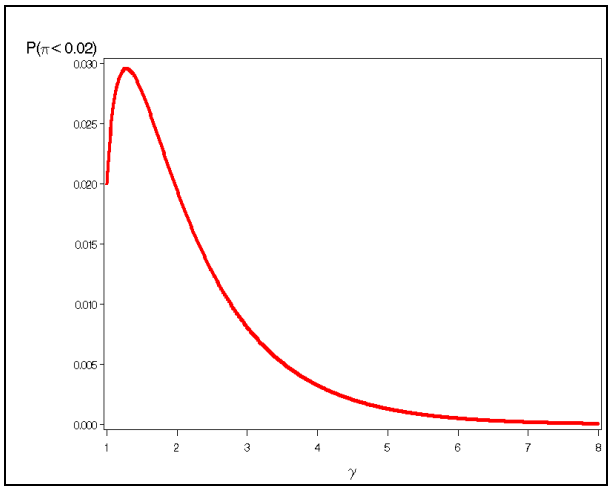
$$\delta = 9\gamma - 8$$

The cumulative distribution function (CDF) for the Beta distribution is given by:

$$F(x) = \frac{\Gamma(\gamma + \delta)}{\Gamma\alpha\Gamma\gamma} \int_0^x t^{\gamma-1}(1-t)^{\delta-1}\,dt$$

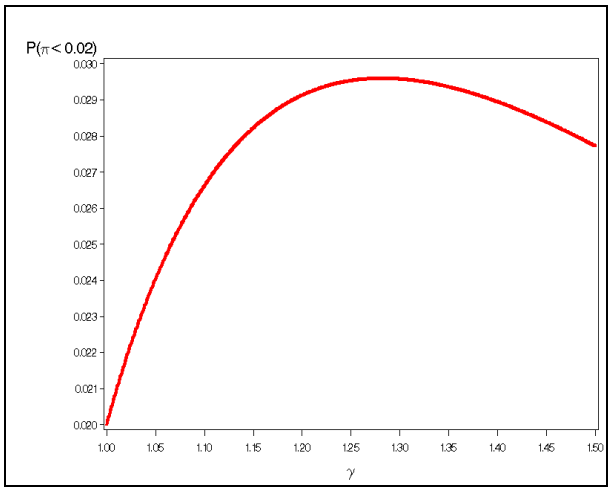
Since it is not straightforward to maximise for this express algebraically, satisfying the relationship between γ and δ given above, a graphical approach was taken in order to identify values of γ and δ that maximise $P(\pi<0.02)$ and $\pi_M = 0.1$ (Figure C.1).

Figure C.1 $P(\pi<0.02)$ for a Beta distribution where $\pi_M = 0.1$



The region of the maximum can be seen more clearly in Figure C.2.

Figure C.2 Maximum $P(\pi>0.02)$ for a Beta distribution where $\pi_M = 0.1$



For convenience the value $\gamma = 1.25$ was chosen, giving the distribution *Beta*(1.25, 3.25) to be used as the prior distribution.

Appendix C.2 Silcock's Property 1 applied to the standardized mortality rate (Page 132)

Consider the ratio of standardized mortality ratios for two populations, *a* and *b*, standardized to population *R*:

$$\frac{SMR_a}{SMR_b} = \frac{\sum_{i=1}^{n_a} p_{ai}\pi_{ai}}{\sum_{i=1}^{n_a} p_{ai}\pi_{Ri}} . \frac{\sum_{i=1}^{n_b} p_{bai}\pi_{Ri}}{\sum_{i=1}^{n_b} p_{bi}\pi_{bi}}$$

Of interest is whether $\alpha \le \frac{\pi_{ai}}{\pi_{bi}} \le b$.

So, given: $u \le \frac{\pi_{ai}}{\pi_{Ri}} \le U$

and $v \le \frac{\pi_{bi}}{\pi_{Ri}} \le V$

it can be seen that since the SMR is a weighted average of these stratum specific ratios we have:

$$u \le SMR_a \le U$$

and $u \le SMR_b \le U$.

This gives:

$$\frac{u}{V} \le \frac{SMR_a}{SMR_b} \le \frac{U}{v}$$

It can be written that $\frac{\pi_{ai}}{\pi_{Ri}} = u + x_i$ and $\frac{\pi_{bi}}{\pi_{Ri}} = V - y_i$, where $x_i \ge 0$ and $y_i \ge 0$.

Now consider the least value of the set $\frac{\pi_{ai}}{\pi_{bi}}$, i.e. α:

$$\alpha = \frac{\pi_{ai}}{\pi_{bi}} = \frac{\pi_{ai}}{\pi_{Ri}} . \frac{\pi_{Ri}}{\pi_{bi}} = \frac{u+x}{V-y} > \frac{u}{V}$$

except in the special case where $x = y = 0$.

This argument can be repeated for the upper limit to give:

$$\frac{u}{V} < \alpha < \beta < \frac{U}{v}$$

Hence, the ratio of the SMRs can vary outside the interval (α,β) contrary to *Property 1*.

The special case $x = y = 0$ is now considered.

First, if $x = y = 0$ for all i, then:

$$\pi_{ai} = k_a \pi_{Ri},\ \pi_{bi} = k_b \pi_{Ri} \text{ and } \frac{\pi_{ai}}{\pi_{bi}} = \frac{k_a}{k_b}.$$

Hence, $$\frac{SMR_a}{SMR_b} = \frac{k_a}{k_b}$$

Second, the case where $x = y = 0$ for some i only requires that $\frac{\pi_{ai}}{\pi_{Ri}}$ takes its lower limit for the same stratum as $\frac{\pi_{bi}}{\pi_{Ri}}$ takes its upper limit, together with a similar condition for the other limits. Such a situation is very unlikely to occur in practice.

Appendix C.3 Derivation of the solution for the lower limit in Equation 5.31 (Page 145)

$$\frac{\bar{d} - \pi_L - \frac{1}{2n}}{\sqrt{\left(\frac{\pi_L[1-\pi_L]}{n}\right)}} = z_{\alpha/2}$$

$$\frac{\left(d - \pi_L - \frac{1}{2n}\right)^2}{\left(\frac{\pi_L[1-\pi_L]}{n}\right)} = z_{\alpha/2}^2$$

$$n\left(\bar{d}^2 - 2\bar{d}\pi_L - \frac{\bar{d}}{n} + \pi_L^2 + \frac{\pi_L}{n} + \frac{1}{4n^2}\right) = z_{\alpha/2}^2\left(\pi_L[1-\pi_L]\right)$$

$$n\bar{d}^2 - 2n\bar{d}\pi_L - \bar{d} + \pi_L^2 + \pi_L + \frac{1}{4n} = z_{\alpha/2}^2\pi_L - z_{\alpha/2}^2\pi_L^2$$

$$\pi_L^2\left(n + z_{\alpha/2}^2\right) + \pi_L\left(1 - 2n\bar{d} - z_{\alpha/2}^2\right) + \left(n\bar{d}^2 - \bar{d} + \frac{1}{4n}\right) = 0$$

The solutions can be found using the quadratic formula:

$$\pi_L = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where: $a = \left(n + z_{\alpha/2}^2\right)$

$$b = \left(1 - 2n\bar{d} - z_{\alpha/2}^2\right)$$

$$c = \left(n\bar{d}^2 - \bar{d} + \frac{1}{4n}\right)$$

Hence:

$$\pi_L = \frac{2n\bar{d} + z_{\alpha/2}^2 - 1 \pm \sqrt{\left(1 - 2n\bar{d} - z_{\alpha/2}^2\right)^2 - 4\left(n + z_{\alpha/2}^2\right)\left(n\bar{d}^2 - \bar{d} + \frac{1}{4n}\right)}}{2\left(n + z_{\alpha/2}^2\right)}$$

$$= \frac{2n\bar{d} + z_{\alpha/2}^2 - 1 \pm \sqrt{1 - 4n\bar{d} + 4n\bar{d}z_{\alpha/2}^2 + 4n^2\bar{d}^2 - 2z_{\alpha/2}^2 + z_{\alpha/2}^4 - 4n^2\bar{d}^2 - 4n\bar{d} + 1 + 4n\bar{d}^2 z_{\alpha/2}^2 - 4\bar{d}z_{\alpha/2}^2 + \frac{z_{\alpha/2}^2}{n}}}{2\left(n + z_{\alpha/2}^2\right)}$$

$$= \frac{2n\bar{d} + z_{\alpha/2}^2 - 1 \pm \sqrt{4n\bar{d}z_{\alpha/2}^2 - 2z_{\alpha/2}^2 + z_{\alpha/2}^4 - 4n\bar{d}^2 z_{\alpha/2}^2 + 4\bar{d}z_{\alpha/2}^2 - \frac{z_{\alpha/2}^2}{n}}}{2\left(n + z_{\alpha/2}^2\right)}$$

$$= \frac{2n\bar{d} + z_{\alpha/2}^2 - 1 \pm \sqrt{z_{\alpha/2}^2\left(4n\bar{d} - 2 + z_{\alpha/2}^2 - 4n\bar{d}^2 + 4\bar{d} - \frac{1}{n}\right)}}{2\left(n + z_{\alpha/2}^2\right)}$$

$$= \frac{2n\bar{d} + z_{\alpha/2}^2 - 1 \pm z_{\alpha/2}\sqrt{z_{\alpha/2}^2 - \left(2 + \frac{1}{n}\right) + 4\bar{d}\left(n - n\bar{d} + 1\right)}}{2\left(n + z_{\alpha/2}^2\right)}$$

$$= \frac{2n\bar{d} + z_{\alpha/2}^2 - 1 \pm z_{\alpha/2}\sqrt{z_{\alpha/2}^2 - \left(2 + \frac{1}{n}\right) + 4\bar{d}\left(n\left[1 - \bar{d}\right] + 1\right)}}{2\left(n + z_{\alpha/2}^2\right)}$$

Appendix C.4 Estimation of the lower 95% confidence limit for the Standardized Mortality Ratio for Unit 9 using the 'full' normal approximation method (Page 247)

The lower limit for a confidence interval for the Standardized Mortality Ratio (SMR) is given by, see (*5.43*):

$$SMR_L = \frac{\pi_L}{\sum_{i=1}^{n} \hat{\pi}_i}$$

where π_L is:

$$\pi_L = \frac{\left(2n\bar{d} + z_{\alpha/2}^2 - 1\right) - z_{\alpha/2}\sqrt{z_{\alpha/2}^2 - \left(2 + \left[1/n\right]\right) + 4\bar{d}\left(n\left[1-\bar{d}\right]+1\right)}}{2\left(n + z_{\alpha/2}^2\right)} \quad \text{given in (5.31)}$$

To calculate the lower limit of a 95% confidence interval for Unit 9, using the complete data for the reduced model (Table 6.31):

$n = 32$

$\bar{d} = 0$

$z_{\alpha/2}^2 = 1.96$

$\sum \hat{\pi}_i = 0.40$

Hence:

$$\pi_L = \frac{\left(0 + 1.96^2 - 1\right) - 1.96\sqrt{1.96^2 - \left(2 + \left[1/32\right]\right) + 0}}{2\left(32 + 1.96\right)} \approx 0.0030$$

and:

$$SMR_L \approx \frac{32 \times 0.0030}{0.40} \approx 0.24$$

Appendix D: SAS PROGRAMMES

This Appendix contains four SAS programmes. The first two are macros written for this thesis: *boot_rank* estimates 95% percentile confidence intervals for the observed rank of each NICU (Page 53) and *weighted*: estimates weighted odds ratios (Page 100). The third programme shown is taken from Luft & Brown (1993) and is their method for calculating exact p-values for the difference between observed and expected mortality (Page 119). The final programme investigates the relationship between gestational age, birth weight and sex using the ratio of the observed birth weight to the gestation-sex specific mean (Page 397).

Appendix D.1 SAS macro *boot_rank* (Page 53)

```
%macro boot_rank;
data all_out;
run;
%do i=1 %to 1000;
      proc surveyselect data=phd.tns method=urs out=sample outhits noprint
                  sampsize=(212 283 38 143 333 378 243 124 35 146 445 196
136 90 124 100);
            strata c_hosp;
      run;

      proc means data=sample nway noprint;
            class c_hosp;
            var died;
            output out=out mean=rate;
      data out;
      set out;
            replicate=&i;
      data all_out;
      set all_out out;
      run;
%end;

proc rank data=all_out out=rank;
      by replicate;
      var rate;
      ranks rank;
run;
proc sort data=rank;
      by c_hosp;
proc univariate data=rank noprint;
      by c_hosp;
      var rank;
      output out=final pctlpre=P_ pctlpts=2.5,50,97.5;
proc print data=final;
run;

%mend boot_rank;
```

Appendix D.2 SAS macro *weighted* (Page 100)

```
%MACRO weighted;

%*****************************************************************;
%* This macro uses weighted logistic regression to estimate odds    ;
%* ratios for unit treatment effects                                ;
%*****************************************************************;

proc printto log="nul:";
data or;
run;

%do j=1 %to 16;

      *Admission in each unit;
      proc means data=tns nway noprint;
      where c_hosp^=&j;
            class c_hosp;
            var died;
            output out=total n=n;

      *Mean number of admissions;
      proc means data=total nway noprint;
            var n;
            output out=meansize mean=meansize;
      data _null_;
      set meansize;
            call symput( 'meansize', put(meansize, best10.) );

      *Calculate weight for each unit;
      data total;
      set total;
            wt=&meansize/n;

      *Apply weight for each observation;
      proc sort data=tns;
            by c_hosp;
      proc sort data=total;
            by c_hosp;
      data weighted;
      merge tns total;
            by c_hosp;
            indicator1=(c_hosp=&j); *Indicator variable;
            if c_hosp=&j then wt=1; *Weight for unit of interest;

      proc logistic data=weighted outest=unit_or covout descending;
      class indicator1 / param=ref ref=first;
            model died = indicator1 gest;
            weight wt;
      run;

      data unit_or;
      set unit_or;
            where _NAME_ in ('died' 'indicator11');
            keep indicator11;
      proc transpose    data=unit_or
                              out=unit_or;
      data unit_or;
```

```
      set unit_or;
            unit=&j;
            odds_ratio=round(exp(col1),.01);
            lower_limit=floor(100*exp(col1-1.96*sqrt(col2)))/100;
            upper_limit= ceil(100*exp(col1+1.96*sqrt(col2)))/100;
            if odds_ratio<1 then
                  p_value=round(2*(probnorm(col1/sqrt(col2))),.001);
            else
                  p_value=round(2*(1-probnorm(col1/sqrt(col2))),.001);
      run;

      data or;
      set or unit_or;
      run;

%end;

proc printto;

proc print data=or noobs;
      where odds_ratio^=.;
      var unit odds_ratio lower_limit upper_limit p_value;
run;

%MEND weighted;
```

Appendix D.3 Luft & Brown (1993) method for calculating exact p-values for the difference between observed and expected mortality (Page119)

```
*     THIS PROGRAM USED A PATIENT-LEVEL DATASET, SORTED BY
      HOSPITAL ID ("HOSPID"). IT HAS A VARIABLE CALLED "DIED", WHICH
      IS A BINARY VARIABLE (1=YES, 0=NO) AND A VARIABLE CALLED
      "P", WHICH CONTAINS THE PREDICTED MORTALITY SCORE.
      IT CALCULATES FOR EACH HOSPITAL 1) PL, THE LOWER TAIL PROBABILITY OF
      OBSERVING D DEATHS OR FEWER AND 2) PU, THE UPPER TAIL PROBABILITY OF
      OBSERVING D DEATHS OR MORE.
      IF THERE ARE 15 OR FEWER DEATHS IN A GIVEN HOSPITAL, IT CALCULATES
      THE EXACT PROBABILITY. OTHERWISE, IT CALCULATES THE PROBABILITY
      USING A NORMAL APPROXIMATION.;
DATA TEMP;
SET __________ (KEEP = HOSPID DIED P);    *FILL IN INPUT DATA SET NAME;
BY HOSPID;
RETAIN SUMDIED SUMPRED SUMPQ SUMPATS LAST0-LAST15;
IF DIED=. OR P=. THEN DELETE;
* WHEN STARTING A NEW HOSPITAL, RESET THE RUNNING TOTALS;
IF FIRST.HOSPID THEN DO;
      SUMDIED=0;
      SUMPRED=0;
      SUMPQ=0;
      SUMPATS=0;
END;
* INCREMENT THE RUNNING TOTALS;
SUMDIED=SUMDIED+DIED;
SUMPRED=SUMPRED+P;
SUMPATS=SUMPATS+1;
Q=1-P;
SUMPQ=SUMPQ+(P*Q);
* ONLY DO THE EXACT PROBABILITY CALCULATIONS WHILE THERE ARE 15 OR;
* FEWER DEATHS IN THIS HOSPITAL;
IF SUMDIED<=15 THEN DO;
      ARRAY CURRENT {*} PROB0 - PROB15;
      ARRAY LAST    {*} LAST0 - LAST15;
      IF FIRST.HOSPID THEN DO;
            CURRENT{1} = Q;
            CURRENT{2} = P;
      END;
      ELSE DO;
            CURRENT{1} = LAST{1}*Q;
            IF SUMPATS <= 15 THEN DO;
                  DO    J=2 TO SUMPATS;
                        I=J-1;
                        CURRENT{J} = (LAST{I}*P) + (LAST{J}*Q);
                  END;
                  * THE VALUE OF J IS NOW SUMPATS+1;
                  CURRENT{J} = LAST{J-1}*P;
            END;
            ELSE DO J = 2 TO 16;
                  I = J - 1;
                  CURRENT{J} = (LAST{I}*P) + (LAST{J}*Q);
            END;
      END;
      DO I=1 TO 16;
```

```
            LAST{I} = CURRENT{I};
      END;
END;
IF LAST.HOSPID THEN DO;
      IF SUMDIED > 0 THEN DO;        * NORMAL APPROXIMATION;
            PL = PROBNORM((SUMDIED +0.5-SUMPRED) / SQRT(SUMPQ));
            PU = 1-PROBNORM((SUMDIED -0.5-SUMPRED) / SQRT(SUMPQ));
      END;
      ELSE DO;                              * EXACT PROBABILITY;
            IF SUMDIED = 0 THEN DO;
                  PU=1;
                  PL = CURRENT{1};
            END;
            ELSE DO;
                  SUMPROBS = 0;
                  INDX = INT(SUMDIED + 0.0001);
                  DO I=1 TO INDX;
                        SUMPROBS = SUMPROBS + CURRENT{I};
                  END;
                  PU = 1 - SUMPROBS;
                  PL = SUMPROBS + CURRENT(INDX+1);
            END;
      END;
      OUTPUT;
END;
KEEP HOSPID SUMPATS SUMDIED SUMPRED PL PU;
PROC PRINT DATA=TEMP;
RUN;
```

Appendix D.4 Relationship between mortality and gestation-sex specific birth weight (Page 397)

This programme is an example of the modelling approach used in §6.3.3 to investigate the relationship between mortality and gestational age, birth weight and sex. The example shown here relates to the use of the ratio of observed birth weight to the estimated gestation-sex specific mean.

```
proc means data=tns nway noprint;    *OBSERVED MEAN BWT FOR GESTATION &
      class gest;                     VARIANCE;
      var bwt;
      output out=var var=var;

proc glm data=var;                   *SMOOTHED BWT VARIANCE FOR GESTATION;
      model var=gest|gest;
      output out=weight p=est_var;

data weight;                         *CALCULATE WEIGHT=1/VAR;
set weight;
      weight=1/est_var;

proc sort data=tns;                  *ADD WEIGHT TO ORIGINAL DATA;
      by gest;
proc sort data=weight;
      by gest;
data tns;
merge tns weight;
      by gest;

proc glm data=tns;                   *SMOOTHED MEAN BIRTH WEIGHT FOR
      model bwt=gest|gest gest|sex;   GESTATION & SEX;
      weight weight;
      output out=ratio p=mean_bwt;

data ratio;                          *CALCULATE RATIO OF BWT TO MEAN;
set ratio;
      ratio=bwt/mean_bwt;
      label gest=' Gestation' bwt='Birthweight (g)';
run;



proc logistic data=ratio descending;
class sex /param=ref ref=first;
      model died=sex|gest|gest|ratio|ratio|ratio/selection=stepwise lackfit
                                                 slentry=.1 slstay=.1;
      output out=ratio p=step_ratio;
run;
```

Appendix E: WINBUGS MODELLING

This Appendix contains some of the WinBUGS programmes used in this thesis and model diagnostics. The first (`model prob`) was used to estimate the in-unit probability of death (§3.4.2). The second model estimated deviation odds ratios (`model or`) (§5.3.2). Finally, further details from the random-effects modelling are given in Appendix E.3.

Appendix E.1 WinBUGS model *PROB*

This model estimated the unadjusted probability of in-unit death for each NICU in §3.4.2. The data were aggregated for each NICU:

```
n[]   died[]
212   21
283   30
38    2
142   6
333   41
378   54
243   29
124   8
35    1
146   5
445   62
196   5
136   3
90    2
124   10
100   6
END
```

The code below illustrates the model with *Uniform*(0,1) specified as the prior distribution for the probability of death for each NICU but other distributions were also used:

```
model prob { for (j in 1:16) {
      died[j] ~ dbin(p[j],n[j])
      p[j] ~ dunif(0.0,1.0)
} }
```

Trace plots showing the value of $\hat{\pi}_j, j \in \{1,3,5\}$, for the first 1,000 iterations (Figure E.1) and plots of the Brooks-Gelman-Rubin statistic (Figure E.2) for five chains are shown below, with *Uniform*(0,1) specified as the prior distribution for the probability of death for each NICU.

Figure E.1 Trace plots for five chains for model PROB

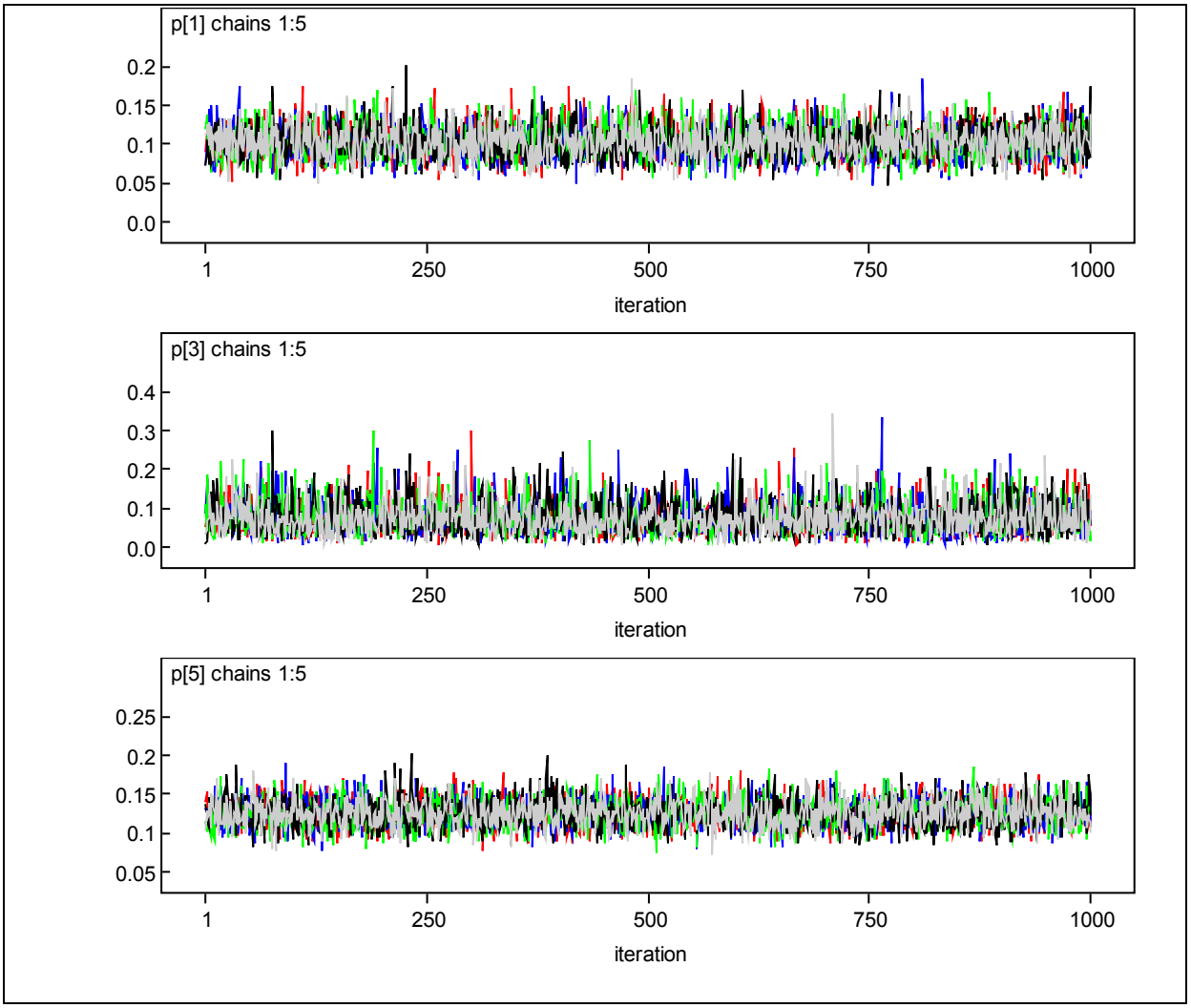
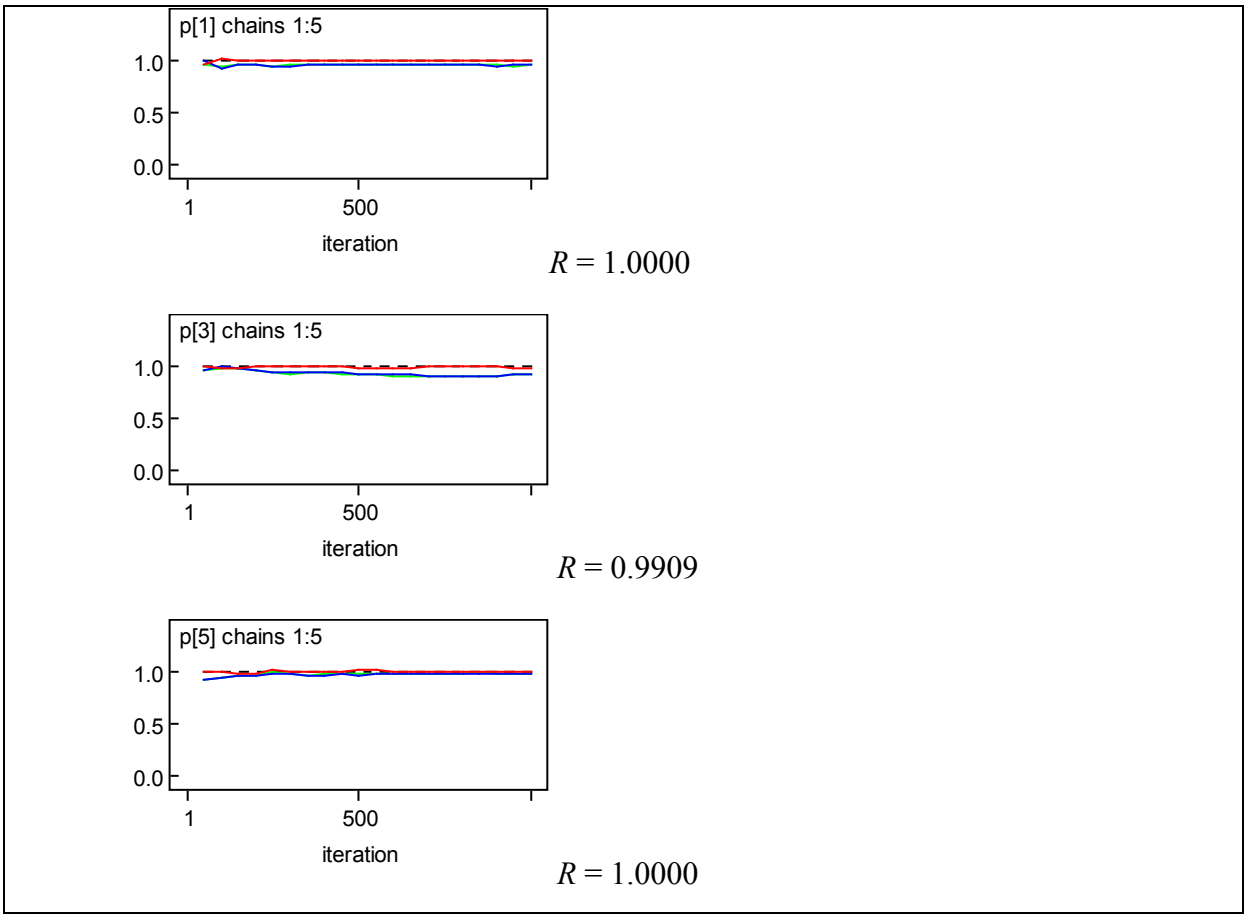


Figure E.2 Plots of Brooks-Gelman-Rubin statistic for model PROB

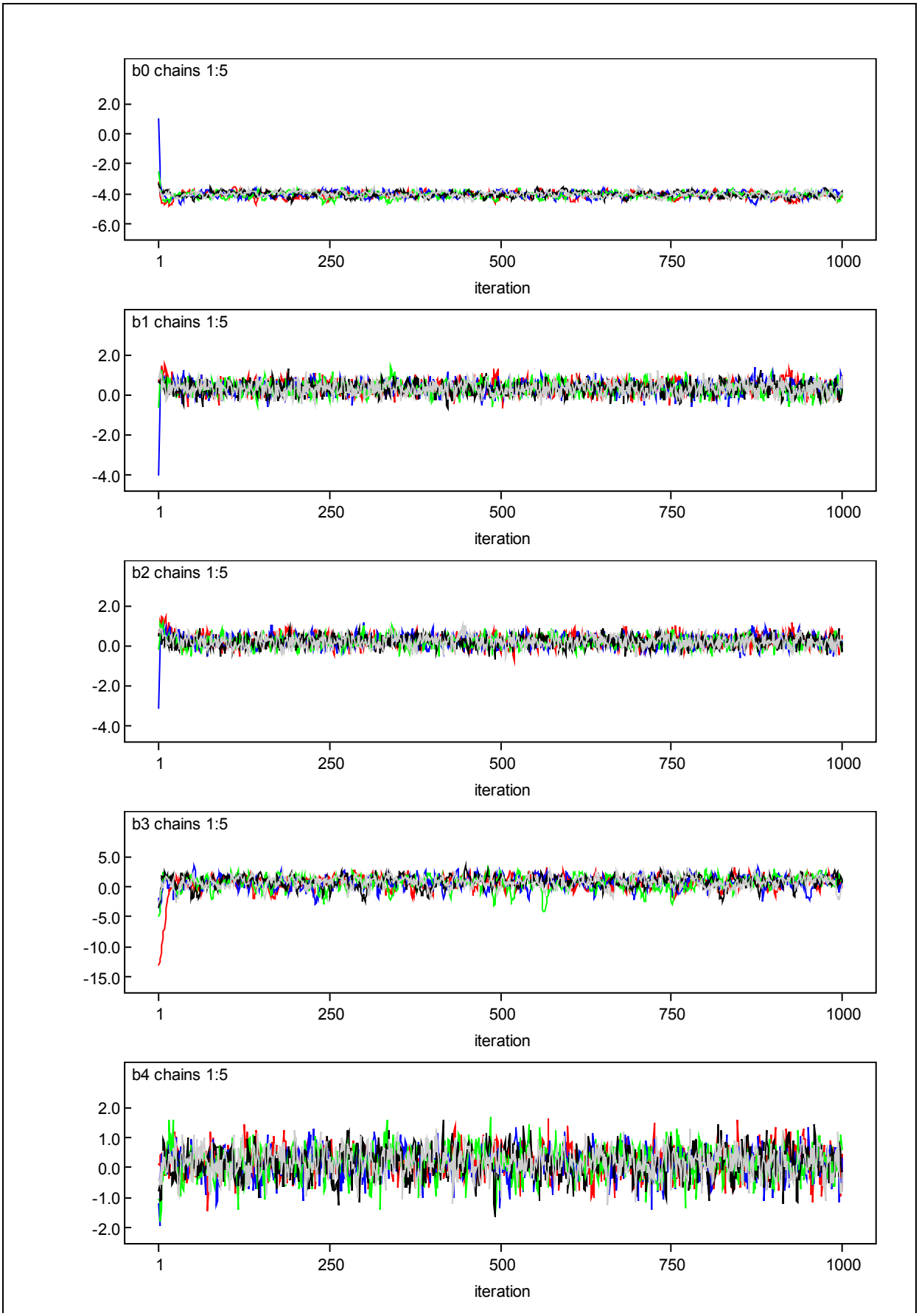


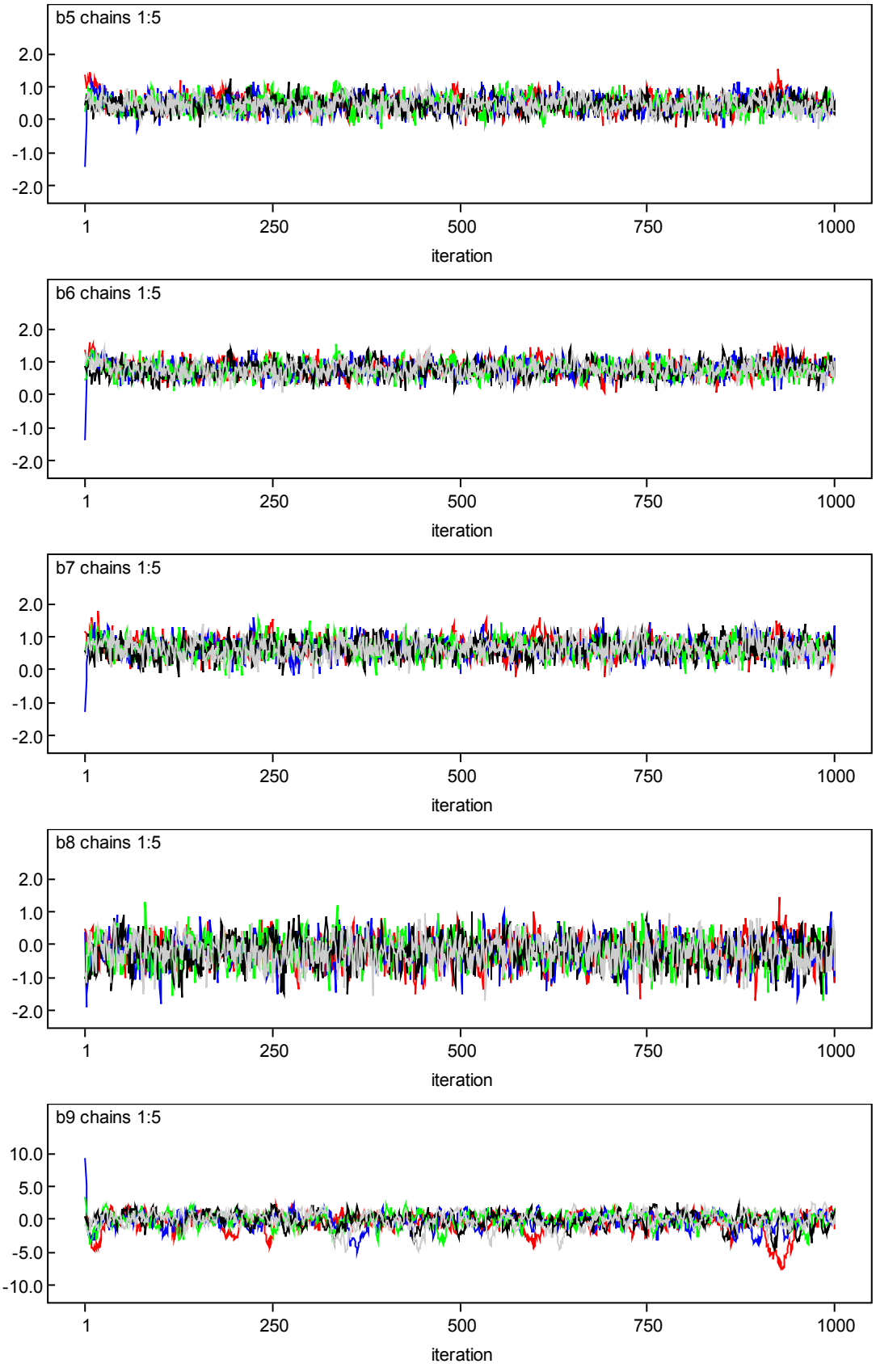
Appendix E.2 WinBUGS model *OR*

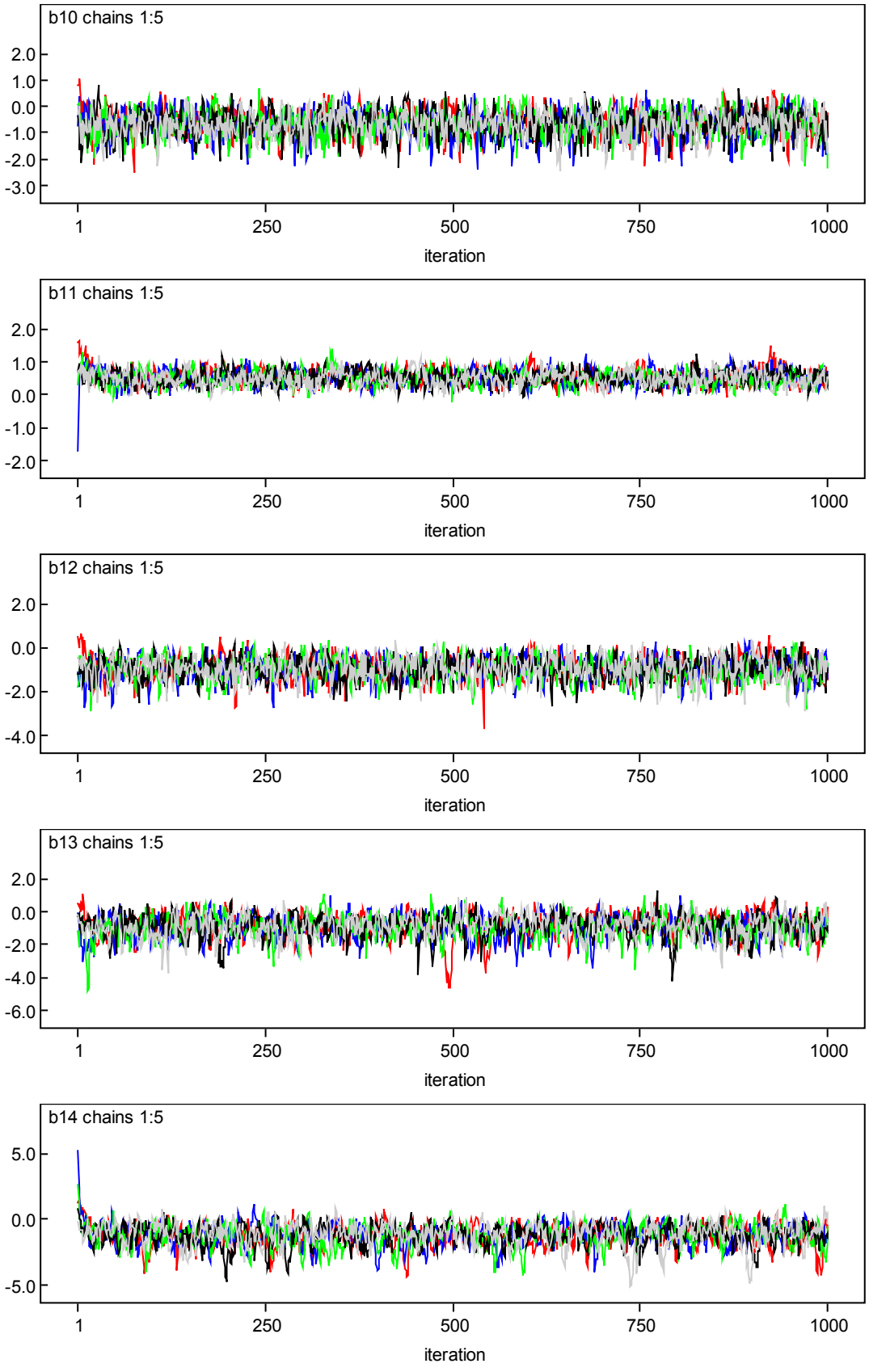
The WinBUGS code reproduced below was used to estimate the deviation odds ratios reported in §5.3.2. Trace plots (Figure E.3) and Brooks-Gelman-Rubin statistic plots (Figure E.4) for five chains over the first 1,000 iterations are also shown for the parameters from the logistic regression model.

```
model or { for (i in 1:3025) {
      died[i] ~ dbern(p[i])

      c_gest[i] <- gest[i]-30       # Centre gestational age

      logit(p[i]) <- b0 + b1*i1[i] + b2*i2[i] + b3*i3[i] + b4*i4[i]
                        + b5*i5[i] + b6*i6[i] + b7*i7[i] + b8*i8[i]
                        + b9*i9[i] + b10*i10[i] + b11*i11[i] + b12*i12[i]
                        + b13*i13[i] + b14*i14[i] + b15*i15[i]
                        + bg*c_gest[i]    }

lor1 <- b1*(16/15)       # Calculate estimated log odds ratio for Unit 1
or1 <- exp(lor1)         # Calculate estimated odds ratio for Unit 1
over1 <- step(or1-2)     # Indicator odds ratio >2 for Unit 1
under1 <- step(0.5-or1)  # Indicator odds ratio <0.5 for Unit 1

Repeat for other units

lor15 <- b15*(16/15)
or15 <- exp(lor15)
over15 <- step(or15-2)
under15 <- step(0.5-or15)

lor16 <- (-16/15)*(b1+b2+b3+b4+b5+b6+b7+b8+b9+b10+b11+b12+b13+b14+b15)
or16 <- exp(lor16)
over16 <- step(or16-2)
under16 <- step(0.5-or16)


# Prior distributions for parameter estimates
      b0 ~ dnorm(0,1.0E-6)
      b1 ~ dnorm(0,1.0E-6)
      b2 ~ dnorm(0,1.0E-6)
      b3 ~ dnorm(0,1.0E-6)
      b4 ~ dnorm(0,1.0E-6)
      b5 ~ dnorm(0,1.0E-6)
      b6 ~ dnorm(0,1.0E-6)
      b7 ~ dnorm(0,1.0E-6)
      b8 ~ dnorm(0,1.0E-6)
      b9 ~ dnorm(0,1.0E-6)
      b10 ~ dnorm(0,1.0E-6)
      b11 ~ dnorm(0,1.0E-6)
      b12 ~ dnorm(0,1.0E-6)
      b13 ~ dnorm(0,1.0E-6)
      b14 ~ dnorm(0,1.0E-6)
      b15 ~ dnorm(0,1.0E-6)
      bg ~ dnorm(0,1.0E-6)    }
```

Figure E.3 Trace plots for five chains for model OR




b5 chains 1:5
2.0
1.0
0.0
-1.0
-2.0
1
250
500
750
1000
iteration
b6 chains 1:5
2.0
1.0
0.0
-1.0
-2.0
1
250
500
750
1000
iteration
b7 chains 1:5
2.0
1.0
0.0
-1.0
-2.0
1
250
500
750
1000
iteration
b8 chains 1:5
2.0
1.0
0.0
-1.0
-2.0
1
250
500
750
1000
iteration
b9 chains 1:5
10.0
5.0
0.0
-5.0
-10.0
1
250
500
750
1000
iteration


b10 chains 1:5
2.0
1.0
0.0
-1.0
-2.0
-3.0
1
250
500
750
1000
iteration
b11 chains 1:5
2.0
1.0
0.0
-1.0
-2.0
1
250
500
750
1000
iteration
b12 chains 1:5
2.0
0.0
-2.0
-4.0
1
250
500
750
1000
iteration
b13 chains 1:5
2.0
0.0
-2.0
-4.0
-6.0
1
250
500
750
1000
iteration
b14 chains 1:5
5.0
0.0
-5.0
1
250
500
750
1000
iteration

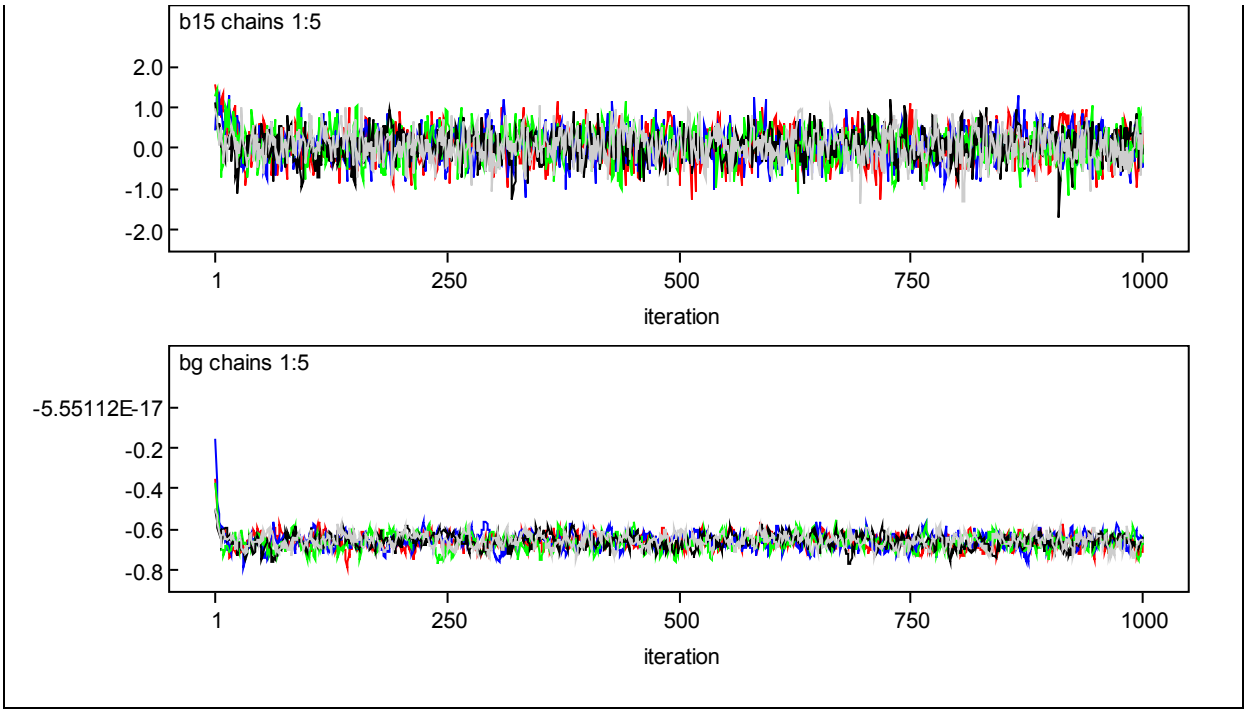
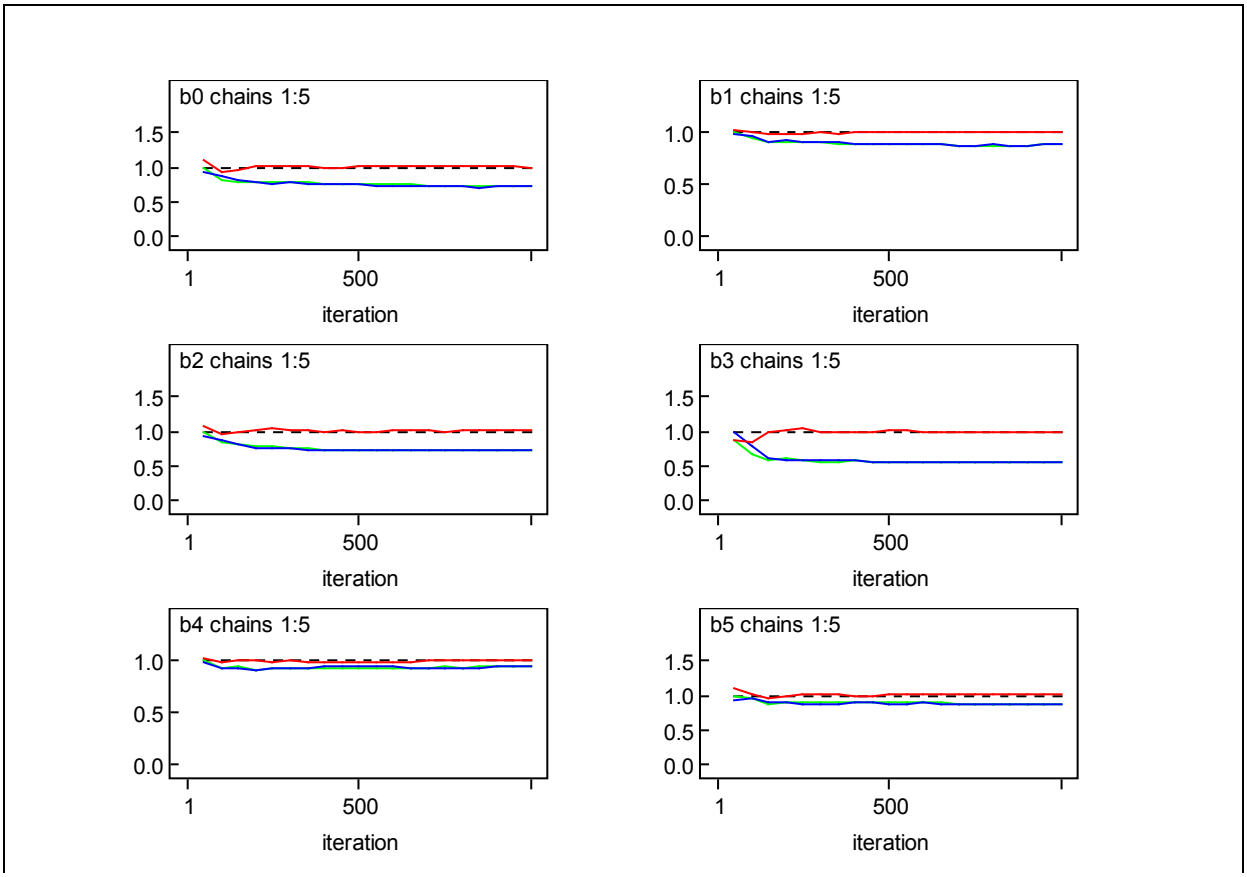
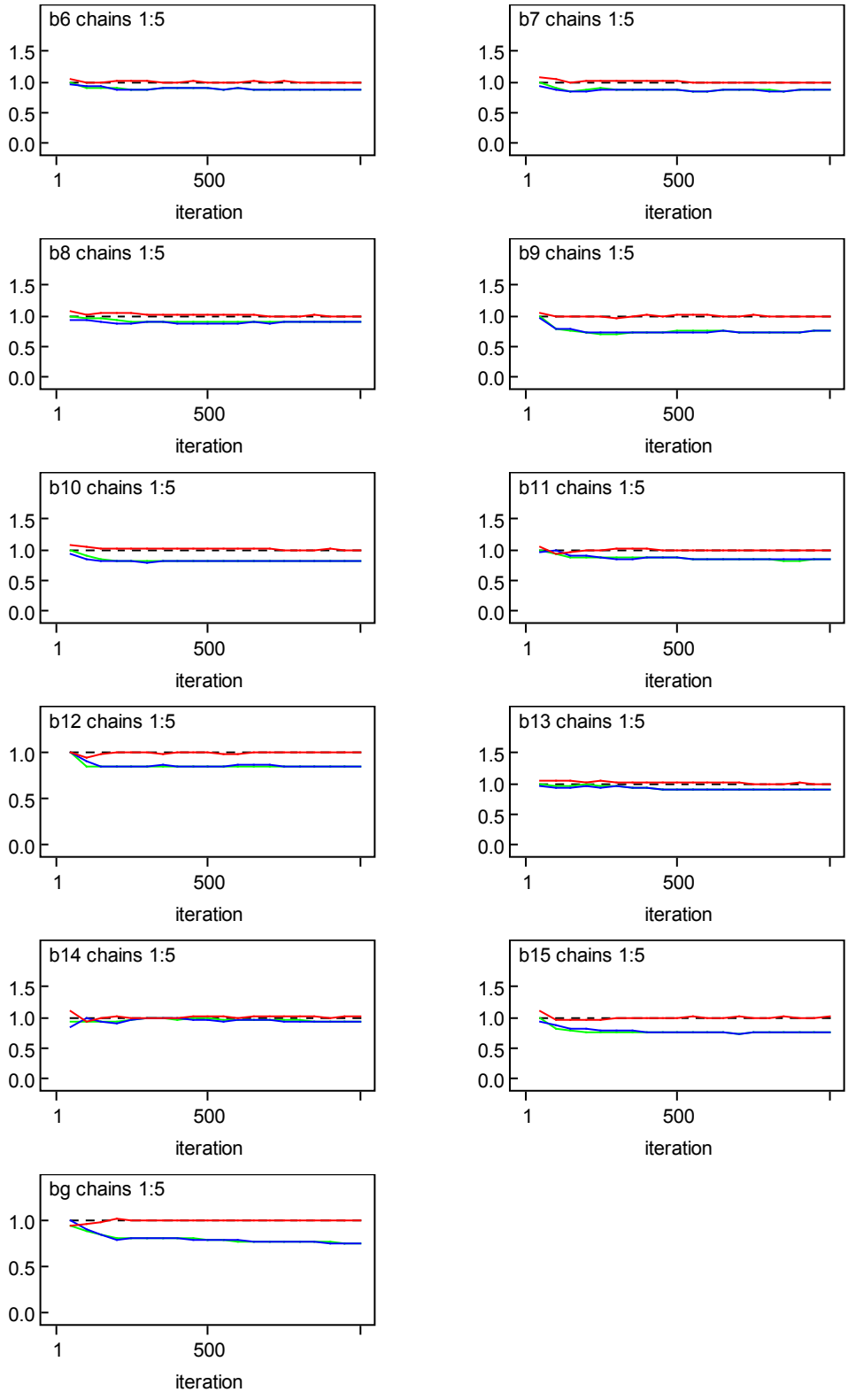


Figure E.4 Plots of Brooks-Gelman-Rubin statistic for model OR




b6 chains 1:5
b7 chains 1:5
b8 chains 1:5
b9 chains 1:5
b10 chains 1:5
b11 chains 1:5
b12 chains 1:5
b13 chains 1:5
b14 chains 1:5
b15 chains 1:5
bg chains 1:5
1.5
1.0
0.5
0.0
1
500
iteration

Appendix E.3 Random Effects Modelling

In this section details are given for the two approaches to model specification in WinBUGS. The differences in sampling properties are illustrated with plots of the Brooks-Gelman-Rubin statistic (R) and of the sampled values (trace plots) over the first 1,000 iterations. Values for the parameters β_0, β_G and σ^2 are shown, together with δ_1 as an example, as all values for δ showed similar properties.

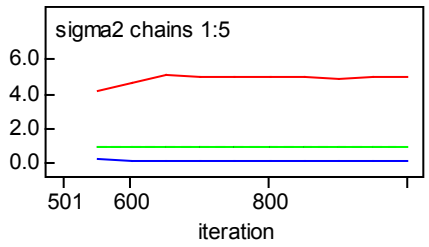
First specification

The parameter β_0 was included in the linear predictor and the mean of the random effect distribution was set to zero:

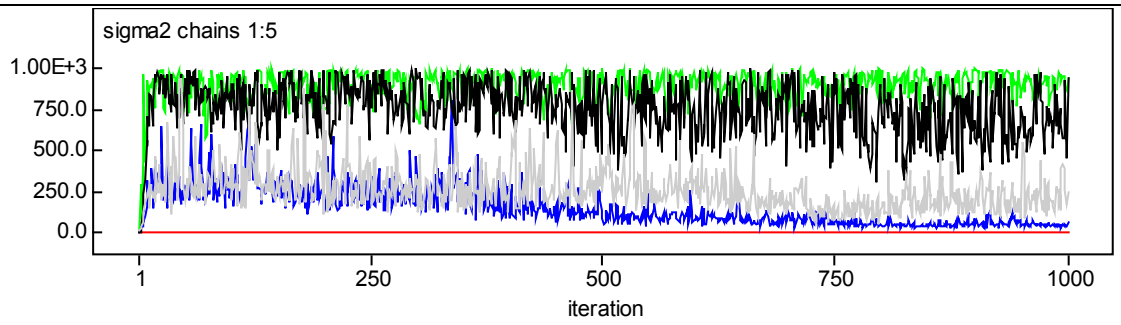
```
{for (i in 1:3025)  {

      c_gest[i] <- gest[i]-30

      died[i]~dbern(p[i])
      logit(p[i])<- b0 + bg*c_gest[i] + delta[c_hosp[i]] }


for (j in 1:16) {

      delta[j]~dnorm(0,tau) }

b0~dnorm(0,1.0E-6)
bg~dnorm(0,1.0E-6)
tau <- 1/sigma2
sigma2~dunif(0,1000)  }
```

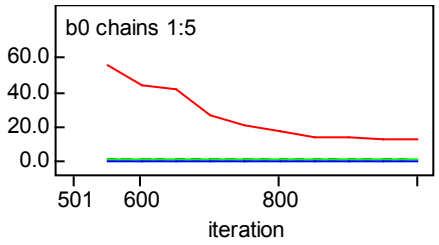
Five chains were run from diverse starting points. However, after 1,000 iterations there was evidence that the chains were not yet sampling from the same sample space, except for β_G (Figure E.5).

Figure E.5 Brooks-Gelman-Rubin statistic and trace plots: first model specification

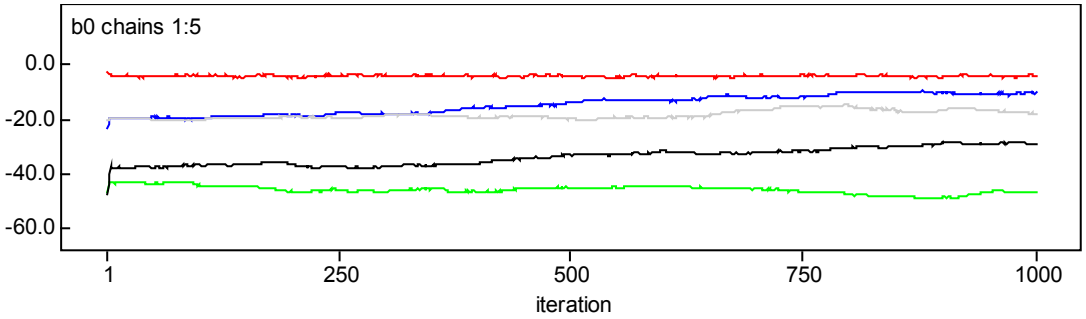


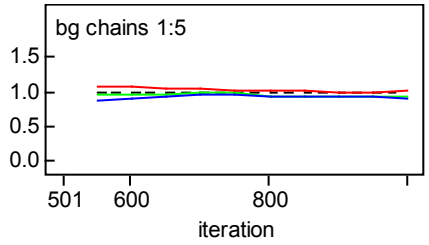
$R_{951\text{-}1000} = 5.01$



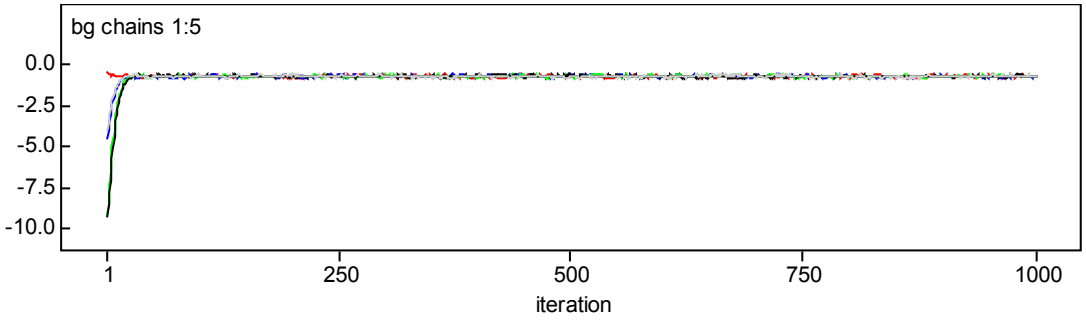


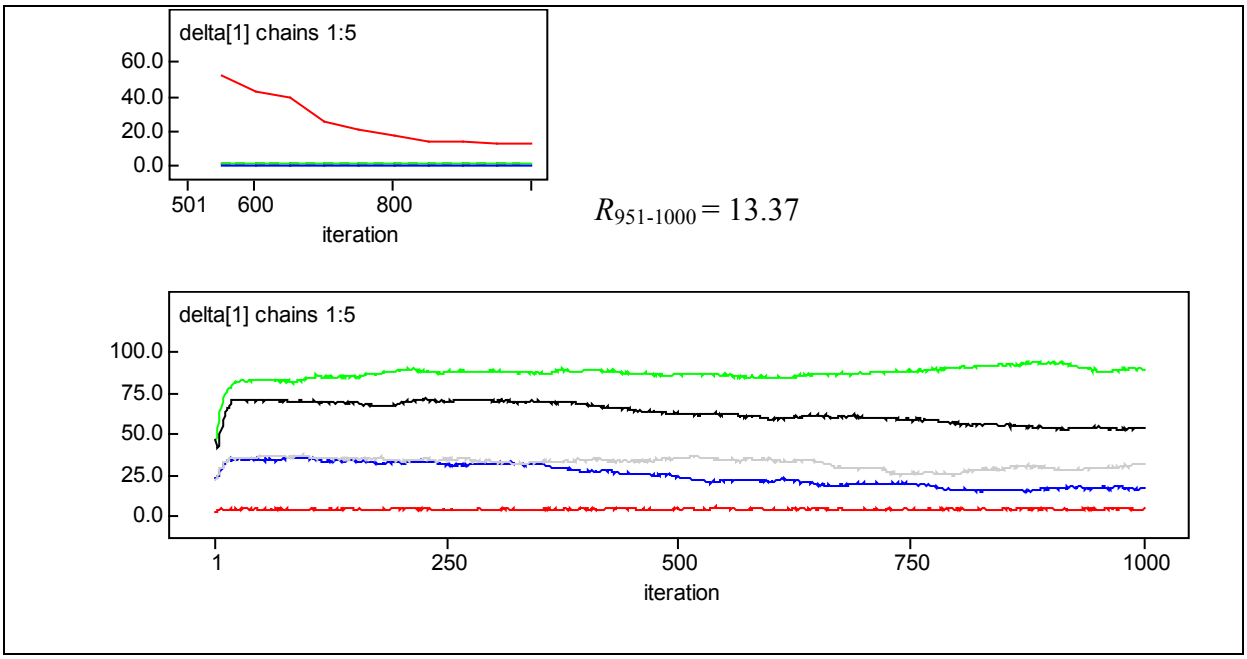
$R_{951\text{-}1000} = 13.44$





$R_{951\text{-}1000} = 1.01$



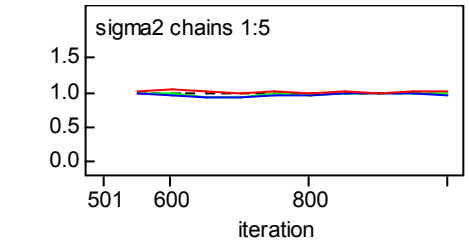


Second specification

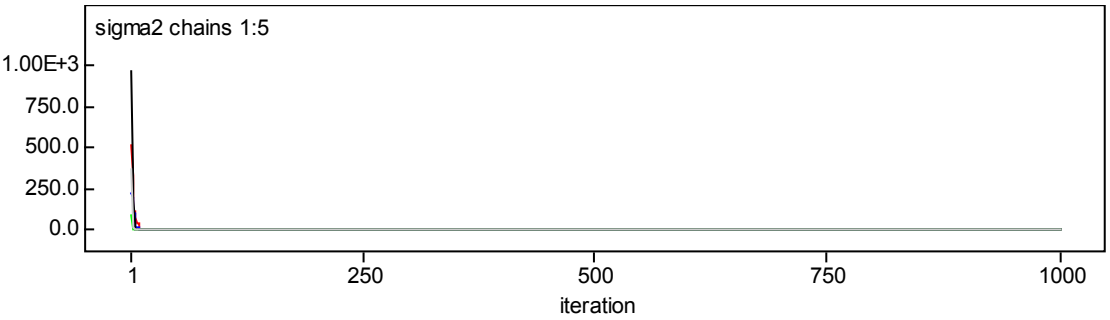
However, if the parameter β_0 was specified as the mean of the random effect, and no intercept included explicitly in the fixed part of the model, the chains showed good mixing (Figure E.6).

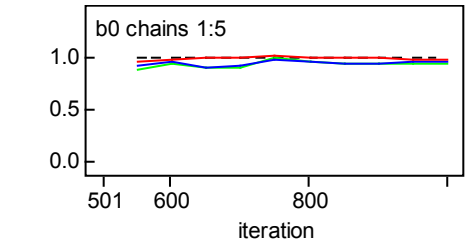
```
{for (i in 1:3025)  {
      died[i]~dbern(p[i])
      c_gest[i] <- gest[i]-30
      logit(p[i])<- bg*c_gest[i] + d[c_hosp[i]] }

for (j in 1:16) {
      d[j]~dnorm(b0,tau)
      delta[j] <- (d[j] - b0) }

b0~dnorm(0,1.0E-6)
bg~dnorm(0,1.0E-6)
tau <- 1/sigma2
sigma2~dunif(0,1000)  }
```

Figure E.6 Brooks-Gelman-Rubin statistic and trace plots: second model specification

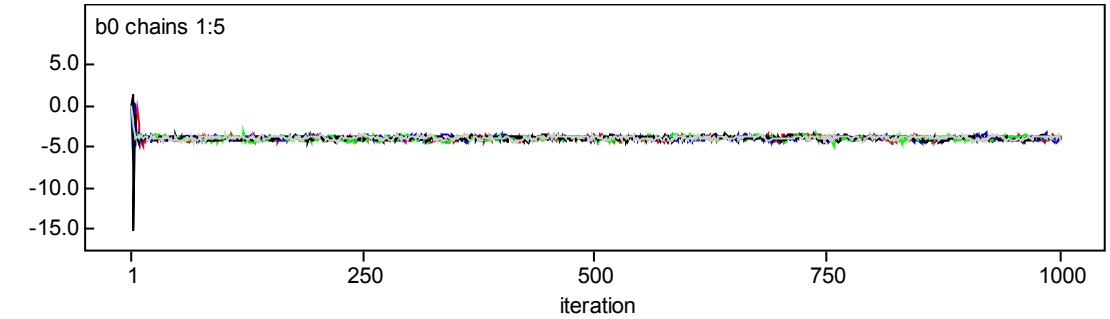


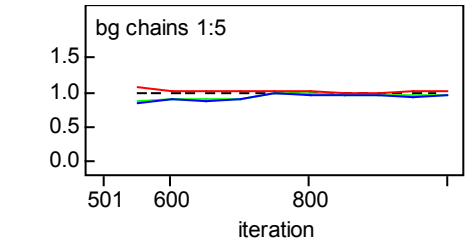
$R_{951\text{-}1000} = 1.01$



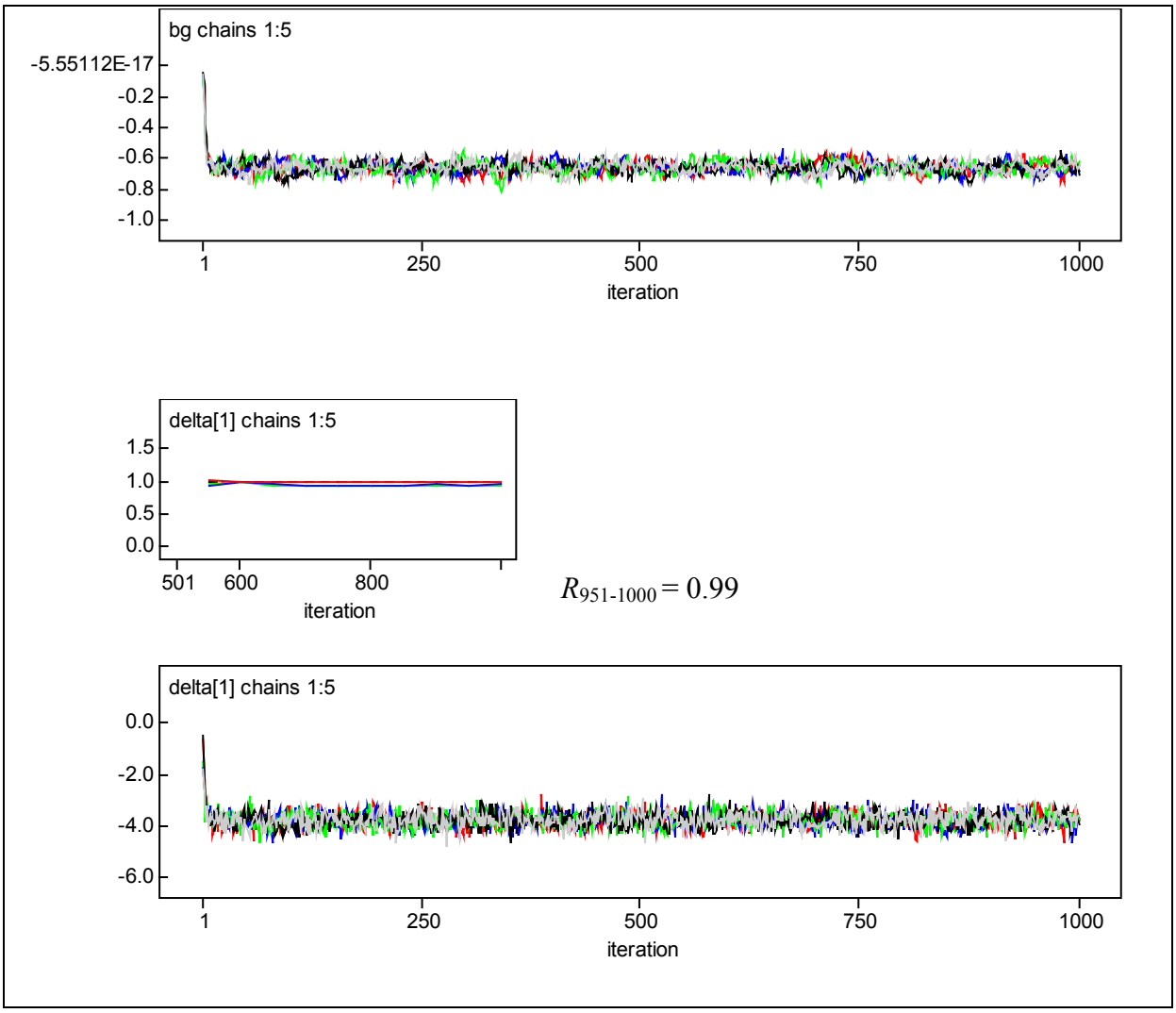


$R_{951\text{-}1000} = 0.99$





$R_{951\text{-}1000} = 1.00$


bg chains 1:5
-5.55112E-17
-0.2
-0.4
-0.6
-0.8
-1.0
1
250
500
750
1000
iteration
delta[1] chains 1:5
1.5
1.0
0.5
0.0
501
600
800
iteration
$R_{951\text{-}1000} = 0.99$
delta[1] chains 1:5
0.0
-2.0
-4.0
-6.0
1
250
500
750
1000
iteration

Appendix F: SIMULATED MORTALITY RATIOS

This appendix contains further details from investigations into possible methods to estimate a confidence interval for a standardized mortality ratio (§5.7). In Appendix F.1 SAS macros and WinBUGS code from the simulation study are shown. The macro *scenario* was used to produce the simulated data sets, and the macros *n_approx*, *ratio_bca*, *h_l*, *z_r* and *full_boot*, together with the given WinBUGS code, were used to estimate the 95% confidence intervals for the SMRs.

Appendix F.2 explores the observed distributions for the SMR. The observed mean, minimum and maximum values of the simulated SMR are shown for each scenario. The observed distributions, together with Normal and Log-Normal probability plots, are plotted for the nine scenarios where the reference population has 1,000 observations (i.e. $\boldsymbol{n_R} = 1000$). The upper and lower limits of the simulated 95% confidence (credible) intervals are compared for various methods where $\boldsymbol{n_j} = 100$ and $\boldsymbol{n_R} = 1000$.

The coverage properties of the methods are given in Appendix F.3. Tables F.2 to F.4 shown the proportion of simulated intervals that did not contain the value of the true SMR. More details are given in Tables F.5 to F.12, where the proportion of intervals falling wholly above and below the true value for the SMR are reported for each method.

Appendix F.1 SAS macros & WinBUGS code

The following SAS macros and the WinBUGS code were used to simulate the data discussed in §5.7 and to estimate the 95% confidence intervals for the SMR.

Simulating the data

```
%MACRO scenario(nsim, Ngroup1, alpha1, beta1, Ngroup2, alpha2, beta2);
%*****************************************************************;
%* This macro creates the data used for the simulation study into   ;
%* methods to estimate confidence intervals for the standardized    ;
%* mortality ratio.  The variables to be specified when calling the ;
%* macro are:                                                       ;
%* nsim: number of simulated data sets (set at 1000 in the thesis)  ;
%* Ngroup1: size of target data set (50, 100 and 200)               ;
%* alpha1:  value for intercept for target data in model            ;
%* beta1:   value for slope for target data set in model            ;
%* Ngroup2: size of reference data set (500, 1000 and 2000)         ;
%* alpha2:  value for intercept for target data in model            ;
%* beta3:   value for slope for target data set in model            ;
%*****************************************************************;
      data score;
            score = 1; output;
            score = 1; output;
            score = 1; output;
            score = 1; output;
            score = 1; output;
            score = 1; output;
            score = 1; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 2; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 3; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 4; output;
            score = 5; output;
            score = 5; output;
            score = 5; output;
```

```
            score = 5; output;
            score = 5; output;
            score = 6; output;
            score = 6; output;
            score = 6; output;
            score = 6; output;
            score = 7; output;
            score = 7; output;
            score = 7; output;
            score = 8; output;
            score = 8; output;
            score = 9; output;
            score = 10; output;

      data data (drop=i j);
      set score;
            do sim=1 to &nsim;                    %*No. simulations;
                  do i=1 to (&Ngroup1/50);        %*Target data;
                        group=1;
                  output;
                  end;
                  do j=1 to (&Ngroup2/50);        %*Reference data;
                  group=2;
                  output;
                  end;
            end;

      proc sort data=data;
            by sim group;

      data data;
      set data;
            if group=1 then do;
                  prob=1/(1+exp(-(&alpha1+&beta1*score)));
                  e_prob=1/(1+exp(-(&alpha2+&beta2*score)));
                  observed=rand('bernoulli',prob);
            end;
            if group=2 then do;
                  prob=1/(1+exp(-(&alpha2+&beta2*score)));
                  observed=rand('bernoulli',prob);
            end;
      run;
%MEND scenario;

%scenario(1000,  50, -2.5, 0.2, 500, -2.5, 0.2);
```

Normal approximation methods

```
%MACRO n_approx(nsim, Ngroup1, alpha1, beta1, Ngroup2, alpha2, beta2);
%*************************************************************;
%* This macro uses the Normal approximation method to estimate      ;
%* confidence intervals for the standardized mortality ratio.       ;
%* The data has previously been created using the 'scenario' macro. ;
%* The variables to be specified when calling the macro are:        ;
%* nsim: number of simulated data sets (set at 1000 in the thesis)  ;
%* Ngroup1: size of target data set (50, 100 and 200)               ;
```

```
%* alpha1:  value for intercept for target data in model            ;
%* beta1:   value for slope for target data set in model            ;
%* Ngroup2: size of reference data set (500, 1000 and 2000)        ;
%* alpha2:  value for intercept for target data in model            ;
%* beta3:   value for slope for target data set in model            ;
%*******************************************************************;
      proc printto log="nul:";

      data data;
      set data;
            if group=1 then m_observed=.;
                  else m_observed=observed;

      %do j=1 %to &nsim;

            proc logistic data=data descending outest=est&j covout noprint;
                  where &j=sim;
                  model m_observed = score;
                  output out=p&j pred=pred;

            data p&j;       %* CALCULATE THE VARIANCE FOR EACH OBSERVED;
            set p&j;
                  where group=1;
                  s2=pred*(1-pred);
                  i=1;
            run;

            proc iml;
                  use work.p&j;
                        read all var{s2} into s2;
                        v=diag(s2);
                        read all var{i} into one;
                  var_o=one`*v*one;
                  create var&j var{var_o};
                  append;
            quit;

            proc means data=p&j sum noprint nway;
                  var pred observed prob e_prob;
                  output out=pred&j sum=/autolabel;

            data count&j;
            merge pred&j var&j;

            proc append base=run_count
                        data=count&j
                        force;

            proc datasets nolist;
                  delete count&j est&j p&j pred&j var&j;
            run;

      %end;

      data run_count;
      set run_count;
            where pred^=.;
            ratio=observed/pred;
            true=prob/e_prob;
            upper=(observed+1.96*sqrt(var_o))/pred;
            under=(upper<true);
```

```
            lower=(observed-1.96*sqrt(var_o))/pred;
            over=(lower>true);
            outside=(under=1 or over=1);

      proc means data=run_count mean n;
            where observed>0;
            var ratio outside under over;
            title "n=&nsim, Group1=&Ngroup1, alpha1=&alpha1, beta1=&beta1,
Group2=&Ngroup2, alpha2=&alpha2, beta2=&beta2";
      run;
      proc printto;
%MEND n_approx;


%n_approx(1000,  50, -2.5, 0.2, 500, -2.5, 0.2);
```

NOTE: This macro uses PROC IML whereas it would probably have been more straightforward to have summed the values of the variable **s2** using a procedure such as PROC MEANS. However, the macro has be written in this way to follow the macros written to implement the methods proposed by Hosmer & Lemeshow and by Zhou & Romano. In practice these three macros can be combined into a single one. They have been show separately here for clarity.

BCa Bootstrap method

```
%MACRO ratio_bca(nsim, Ngroup1, alpha1, beta1, Ngroup2, alpha2, beta2);
%*****************************************************************;
%* This macro uses the bca bootstrap method to estimate          ;
%* confidence intervals for the standardized mortality ratio.    ;
%* It uses two macros available from the SAS website: boot & bootci ;
%* The data has previously been created using the 'scenario' macro. ;
%* The variables to be specified when calling the macro are:     ;
%* nsim: number of simulated data sets (set at 1000 in the thesis)  ;
%* Ngroup1: size of target data set (50, 100 and 200)            ;
%* alpha1:  value for intercept for target data in model         ;
%* beta1:   value for slope for target data set in model         ;
%* Ngroup2: size of reference data set (500, 1000 and 2000)      ;
%* alpha2:  value for intercept for target data in model         ;
%* beta3:   value for slope for target data set in model         ;
%*****************************************************************;
      proc printto log="nul:";
      %include 'k:work\phd\ratio ci\output\bootstrap\sas macros.sas';
      data data;
      set data;
            if group=1 then m_observed=.;
                  else m_observed=observed;

      %do j=1 %to 1000;
            proc logistic data=data descending noprint;
            where sim=&j;
                  model m_observed = score;
                  output out=pred pred=pred;
```

```
            data analyze;
            set pred;
                  where group=1;
            run;

            %MACRO ANALYZE(data= , out= );
                  proc means data=&data nway noprint;
                        var observed pred;
                        output out=totals sum=;
                        %bystmt;
                  data totals;
                  set totals;
                        ratio=observed/pred;
                        %bystmt;
                        keep ratio _sample_;
                  data &out;
                  set totals;
                  run;
            %mend;

            %boot(data=analyze,samples=2000,random=123, stat=ratio,
            print=0, chart=0)

            %bootci(bca,alpha=.05, print=0);

            data bootci;
            set bootci;
                  where name='ratio';
            data ratio;
            set ratio work.bootci;
                  keep value alcl aucl;
            run;
      %end;
      data ratio;
      set ratio;
            under_bca=(aucl<1);
            over_bca=(alcl>1) ;
            outside_bca=(under_bca=1 or over_bca>=1);
      proc means data=ratio mean n;
            var value outside_bca under_bca over_bca;
            title "n=&nsim, Group1=&Ngroup1, alpha1=&alpha1, beta1=&beta1,
Group2=&Ngroup2, alpha2=&alpha2, beta2=&beta2";
      run;

      proc printto;
%mend ratio_bca;

%ratio_bca(1000,  50, -2.5, 0.2, 500, -2.5, 0.2);
```

NOTE: This macro uses two other macros (***boot*** & ***bootci***) available from the SAS Institute website (www.sas.com).

Method proposed by Hosmer & Lemeshow (1995)

```
%MACRO h_l(nsim, Ngroup1, alpha1, beta1, Ngroup2, alpha2, beta2);
%*****************************************************************;
%* This macro uses the Hosmer & Lemeshow's method to estimate    ;
%* confidence intervals for the standardised mortality ratio.    ;
%* The data has previously been created using the 'scenario' macro. ;
%* The variables to be specified when calling the macro are:     ;
%* nsim: number of simulated data sets (set at 1000 in the thesis)  ;
%* Ngroup1: size of target data set (50, 100 and 200)            ;
%* alpha1:  value for intercept for target data in model         ;
%* beta1:   value for slope for target data set in model         ;
%* Ngroup2: size of reference data set (500, 1000 and 2000)      ;
%* alpha2:  value for intercept for target data in model         ;
%* beta3:   value for slope for target data set in model         ;
%*****************************************************************;
      proc printto log="nul:";

      data data;
      set data;
            if group=1 then m_observed=.;
                  else m_observed=observed;

      %do j=1 %to &nsim;

            proc logistic data=data descending outest=est&j covout noprint;
                  where &j=sim;
                  model m_observed = score;
                  output out=p&j pred=pred;

            data p&j;       %* CALCULATE THE VARIANCE FOR EACH OBSERVED;
            set p&j;
                  where group=1;
                  s2=pred*(1-pred);
                  i=1;
            run;

            proc iml;
                  use work.p&j;
                        read all var{s2} into s2;
                        v=diag(s2);
                        read all var{i} into one;
                        read all var{i score} into x;
                  use work.est&j;
                        read all var{Intercept,score} into s

      where(_NAME_=*{"Intercept","score"});
                        var_o=one`*v*one;
                  var_pi=one`*v*x*s*x`*v*one;
                  create var&j var{var_o var_pi};
                  append;
            quit;

            proc means data=p&j sum noprint nway;
                  var pred observed prob e_prob;
                  output out=pred&j sum=/autolabel;

            data count&j;
            merge pred&j var&j;

            proc append base=run_count
```

```
                        data=count&j
                        force;

            proc datasets nolist;
                  delete count&j est&j p&j pred&j var&j;
            run;

      %end;

      data run_count;
      set run_count;
            where pred^=.;
            ratio=observed/pred;
            true=prob/e_prob;
            var_lnR=(var_o/(observed**2))+(var_pi/(pred**2));
            upper_HL=exp(log(ratio)+1.96*sqrt(var_lnR));
            under_HL=(upper_HL<true);
            lower_HL=exp(log(ratio)-1.96*sqrt(var_lnR));
            over_HL=(lower_HL>true);
            outside_HL=(under_HL=1 or over_HL=1);

      proc means data=run_count mean n;
            where observed>0;
            var ratio outside_HL under_HL over_HL;
            title "n=&nsim, Group1=&Ngroup1, alpha1=&alpha1, beta1=&beta1,
Group2=&Ngroup2, alpha2=&alpha2, beta2=&beta2";
      run;
      proc printto;
%MEND h_l;


%h_l(1000,  50, -2.5, 0.2, 500, -2.5, 0.2);
```

Method proposed by Zhou & Romano (1997)

```
%MACRO z_r(nsim, Ngroup1, alpha1, beta1, Ngroup2, alpha2, beta2);
%*****************************************************************;
%* This macro uses the Zhou & Romano's method to estimate        ;
%* confidence intervals for the standardised mortality ratio.    ;
%* The data has previously been created using the 'scenario' macro. ;
%* The variables to be specified when calling the macro are:     ;
%* nsim: number of simulated data sets (set at 1000 in the thesis)  ;
%* Ngroup1: size of target data set (50, 100 and 200)            ;
%* alpha1:  value for intercept for target data in model         ;
%* beta1:   value for slope for target data set in model         ;
%* Ngroup2: size of reference data set (500, 1000 and 2000)      ;
%* alpha2:  value for intercept for target data in model         ;
%* beta3:   value for slope for target data set in model         ;
%*****************************************************************;
      proc printto log="nul:";

      data data;
      set data;
            if group=1 then m_observed=.;
                  else m_observed=observed;
```

```
%do j=1 %to &nsim;

      proc logistic data=data descending outest=est&j covout noprint;
            where &j=sim;
            model m_observed = score;
            output out=p&j pred=pred;

      data p&j;       %* CALCULATE THE VARIANCE FOR EACH OBSERVED;
      set p&j;
            where group=1;
            s2=pred*(1-pred);
            i=1;
      run;

      proc iml;
            use work.p&j;
                  read all var{s2} into s2;
                  v=diag(s2);
                  read all var{i} into one;
                  read all var{i score} into x;
            use work.est&j;
                  read all var{Intercept,score} into s

where(_NAME_=*{"Intercept","score"});
                  var_o=one`*v*one;
            var_pi=one`*v*x*s*x`*v*one;
            create var&j var{var_o var_pi};
            append;
      quit;

      proc means data=p&j sum noprint nway;
            var pred observed prob e_prob;
            output out=pred&j sum=/autolabel;

      data count&j;
      merge pred&j var&j;

      proc append base=run_count
                  data=count&j
                  force;

      proc datasets nolist;
            delete count&j est&j p&j pred&j var&j;
      run;

%end;

data run_count;
set run_count;
      where pred^=.;
      ratio=observed/pred;
      true=prob/e_prob;
      var_R=(ratio**2)*(var_o/(observed**2))+(var_pi/(pred**2));
      upper_ZR=ratio+1.96*sqrt(var_R);
      under_ZR=(upper_ZR<true);
      lower_ZR=ratio-1.96*sqrt(var_R);
      over_ZR=(lower_ZR>true);
      outside_ZR=(under_ZR=1 or over_ZR>=1);

proc means data=run_count mean n;
```

```
            where observed>0;
            var ratio outside_ZR under_ZR over_ZR;
            title "n=&nsim, Group1=&Ngroup1, alpha1=&alpha1, beta1=&beta1,
Group2=&Ngroup2, alpha2=&alpha2, beta2=&beta2";
      run;
      proc printto;
%MEND z_r;


%z_r(1000,  50, -2.5, 0.2, 500, -2.5, 0.2);
```

Bootstrap method

```
%MACRO full_boot(nsim, Ngroup1, alpha1, beta1, Ngroup2, alpha2, beta2);;
      proc printto log="nul:";
      data limits;
      data data;
      set data;
            if group=1 then m_observed=.;
                  else m_observed=observed;
      proc sort data=data;
            by group;
      run;

      %do sim=1 %to 10;
            proc surveyselect data=data method=urs sampsize=(50 1000)
                              out=boot outhits rep=10;
                  where sim=&sim;
                  strata group;
            proc sort data=boot;
                  by replicate;
            proc logistic data=boot descending noprint;
            class replicate / param=glm ref=last;
                  model m_observed = score;
                  by replicate;
                  output out=pred pred=pred;
            proc means data=pred nway noprint;
                  where group=1;
                  class replicate;
                  var observed pred;
                  output out=totals sum=;
            data totals;
            set totals;
                  ratio&sim=observed/pred;
                  keep ratio&sim replicate;
            proc univariate data=totals noprint;
                  var ratio&sim;
                  output out=final pctlpre=P_ pctlpts=2.5,97.5;
            data limits;
            set final limits;
            run;
      %end;

      data limits;
      set limits;
      where P_2_5>.;
```

```
            under_boot=(P_97_5<1);
            over_boot=(P_2_5>1) ;
            outside_boot=(under_boot=1 or over_boot>=1);
      proc means data=limits mean n;
            var outside_boot under_boot over_boot;
            title "n=&nsim, Group1=&Ngroup1, alpha1=&alpha1, beta1=&beta1,
Group2=&Ngroup2, alpha2=&alpha2, beta2=&beta2";
      run;
      proc printto;
      %mend full_boot;


%full_boot(1000,  50, -2.5, 0.2, 500, -2.5, 0.2);
```

Bayesian method

```
model ci {
      for (j in 1:1000) {
            for (i in 11:20) {                  # Reference data

# Estimate model parameters from reference data
                  obs[j,i] ~ dbin(p[j,i],total[j,i])
                  logit(p[j,i]) <- b0[j] + b1[j]*score[j,i] }


            for (i in 1:10)   {           # Data from unit of interest

# Calculate expected 'p' using b0 & b1 estimated above
                  logit(tpp[j,i]) <- b0[j] + b1[j]*score[j,i]

# Number of expected events at each value of 'score' (p * n)
                  pp[j,i] <- tpp[j,i] * total[j,i]


# Estimate model parameters b2 &p3 from reference data
                  obs[j,i] ~ dbin(op[j,i],total[j,i])
                  logit(op[j,i]) <- b2[j] + b3[j]*score[j,i]

# 'Observed' using estimated prob from second logistic model
                  n.obs[j,i] <- op[j,i]*total[j,i]    }


                  sum.pp[j] <- sum(pp[j,])      # Sum of predicted
                  s.ob[j] <- sum(n.obs[j,])     # Sum of observed
                  ratio[j] <- s.ob[j]/sum.pp[j] # Ratio of interest
                  b0[j] ~ dnorm(0,1.0E-6)
                  b1[j] ~ dnorm(0,1.0E-6)
                  b2[j] ~ dnorm(0,1.0E-6)
                  b3[j] ~ dnorm(0,1.0E-6)  } }
```

Appendix F.2 Distribution of simulated standardized mortality ratios

The observed mean, minimum and maximum values of the simulated SMR are shown for each scenario (Table F.1). The size of the reference population is given by n_R and the size of the unit of interest is n_j.

Table F.1 Distribution of simulated standardized mortality ratios

		n_R					
		500		1000		2000	
n_j	Group	Mean	(Range)	Mean	(Range)	Mean	(Range)
Ratio = 1							
50	**1**	0.16	(0.02 – 0.40)	0.16	(0.02 – 0.34)	0.16	(0.02 – 0.32)
	2	0.16	(0.11 – 0.22)	0.16	(0.12 – 0.20)	0.16	(0.14 – 0.19)
100	**1**	0.16	(0.07 – 0.28)	0.16	(0.05 – 0.29)	0.16	(0.06 – 0.29)
	2	0.16	(0.11 – 0.21)	0.16	(0.13 – 0.20)	0.16	(0.13 – 0.19)
200	**1**	0.16	(0.09 – 0.24)	0.16	(0.08 – 0.23)	0.16	(0.09 – 0.27)
	2	0.16	(0.11 – 0.21)	0.16	(0.12 – 0.20)	0.16	(0.13 – 0.18)
Ratio ≈ 1.37							
50	**1**	0.22	(0.06 – 0.40)	0.22	(0.06 – 0.48)	0.22	(0.06 – 0.40)
	2	0.16	(0.11 – 0.21)	0.16	(0.13 – 0.20)	0.16	(0.14 – 0.19)
100	**1**	0.22	(0.10 – 0.35)	0.22	(0.10 – 0.37)	0.22	(0.10 – 0.36)
	2	0.16	(0.11 – 0.21)	0.16	(0.12 – 0.20)	0.16	(0.13 – 0.18)
200	**1**	0.22	(0.13 – 0.33)	0.22	(0.13 – 0.31)	0.22	(0.14 – 0.30)
	2	0.16	(0.11 – 0.22)	0.16	(0.12 – 0.20)	0.16	(0.14 – 0.18)
Ratio ≈ 0.71							
50	**1**	0.11	(0.02 – 0.28)	0.11	(0.00 [i] – 0.30)	0.11	(0.00 [ii] – 0.22)
	2	0.16	(0.11 – 0.23)	0.16	(0.12 – 0.20)	0.16	(0.13 – 0.19)
100	**1**	0.11	(0.02 – 0.22)	0.11	(0.03 – 0.22)	0.11	(0.03 – 0.24)
	2	0.16	(0.10 – 0.21)	0.16	(0.13 – 0.19)	0.16	(0.13 – 0.18)
200	**1**	0.11	(0.05 – 0.18)	0.11	(0.05 – 0.19)	0.11	(0.06 – 0.20)
	2	0.16	(0.11 – 0.22)	0.16	(0.13 – 0.20)	0.16	(0.13 – 0.19)

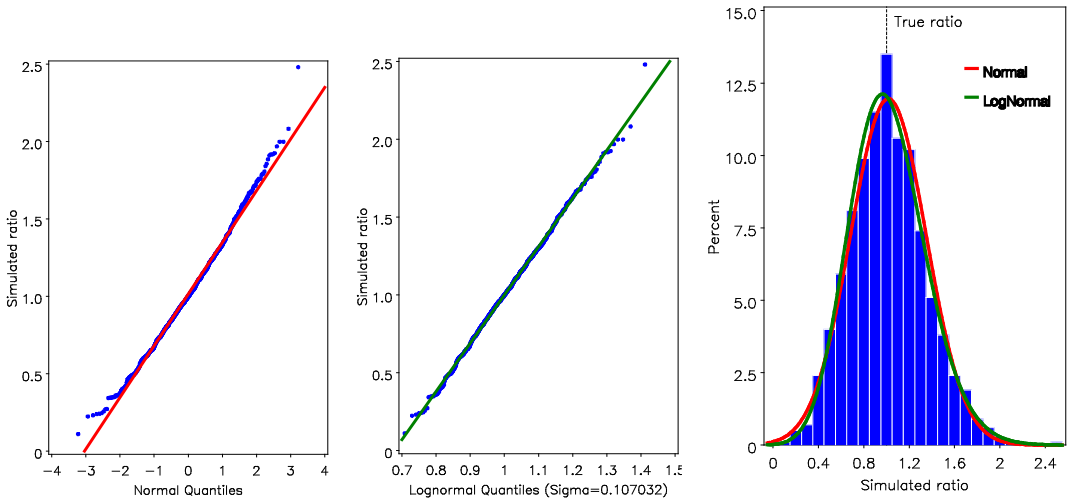
[i] Five data sets with zero observed events [ii] Two data sets with zero observed events

The observed distributions are plotted for the nine scenarios where n_R = 1000, together with Normal and Log-Normal probability plots (Figure F.1).

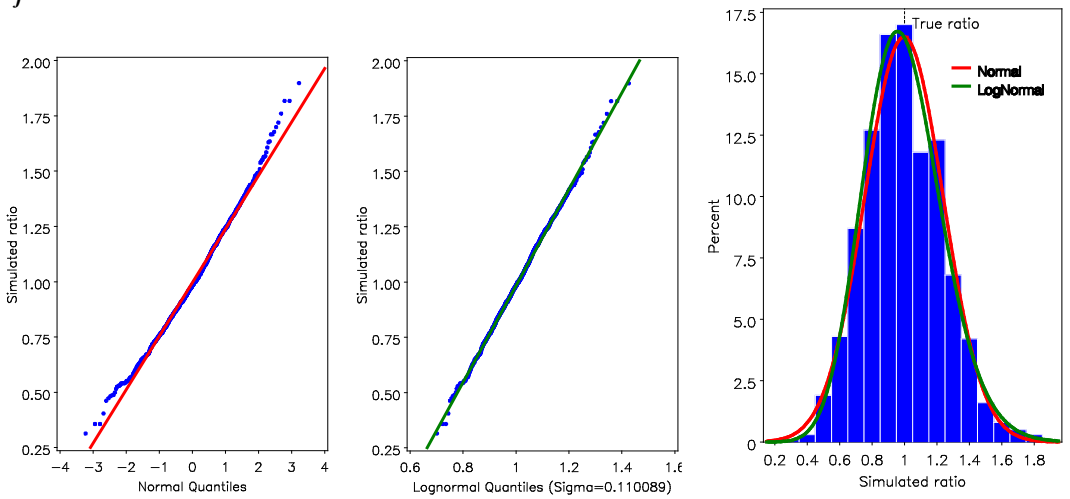
Figure F.1 Simulated Standardized Mortality Ratios

SMR = 1, n_R = 1000

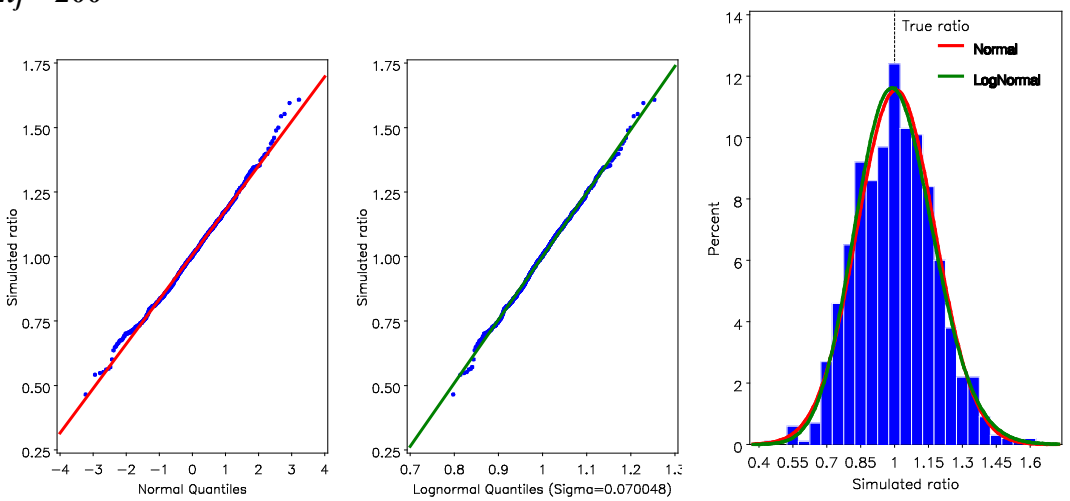
n_j = 50



n_j = 100

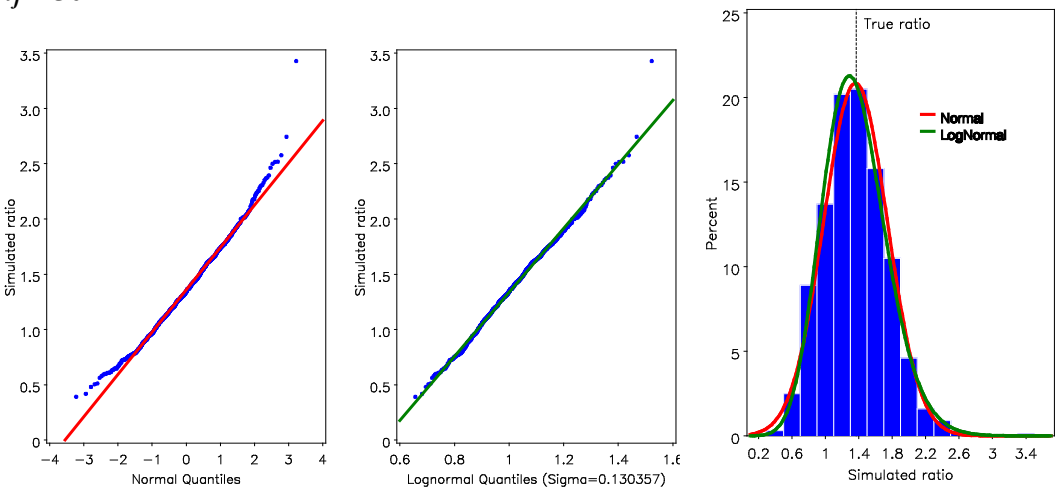


n_j = 200

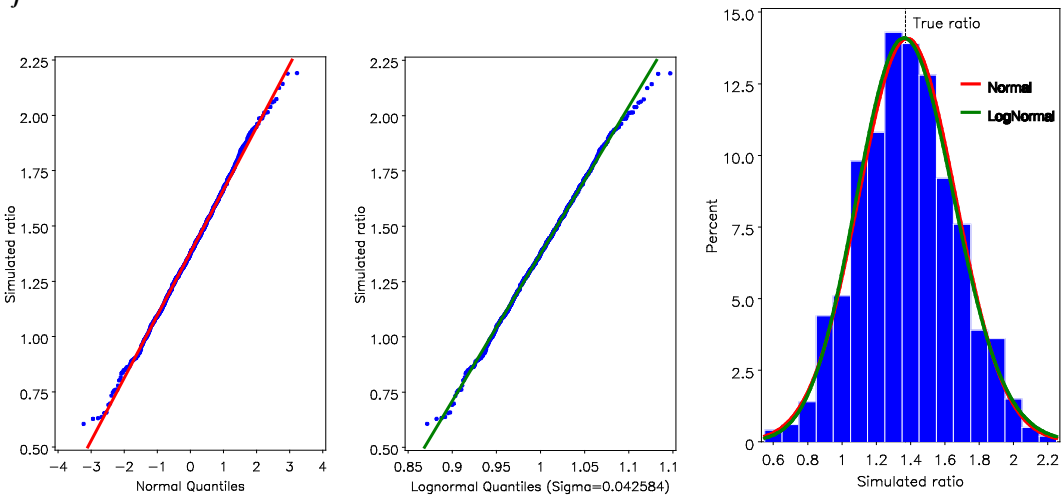


SMR ≈ 1.37, $n_R = 1000$

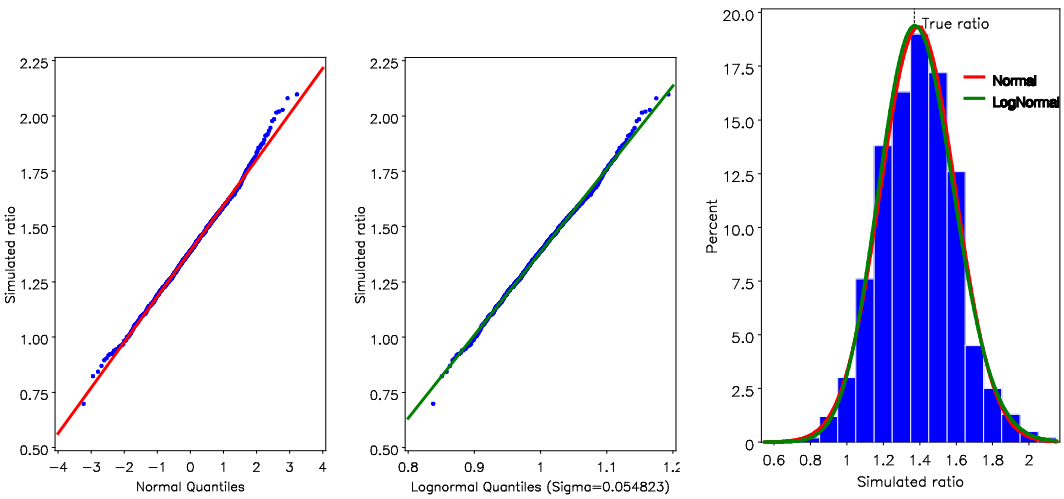
$n_j = 50$



$n_j = 100$

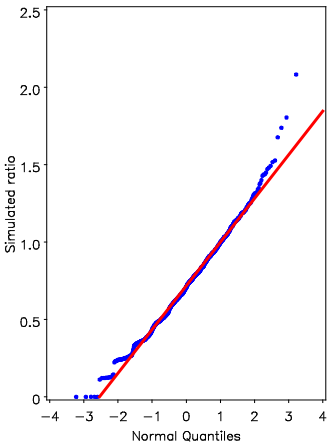


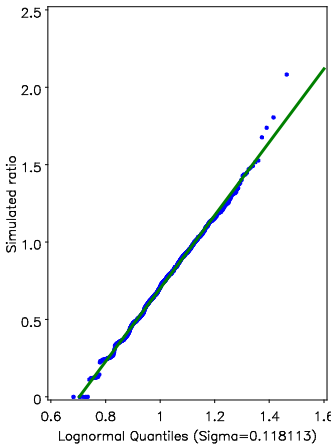
$n_j = 200$

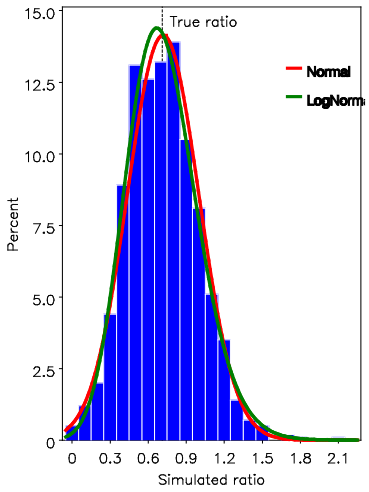


SMR ≈ 0.71, $n_R = 1000$

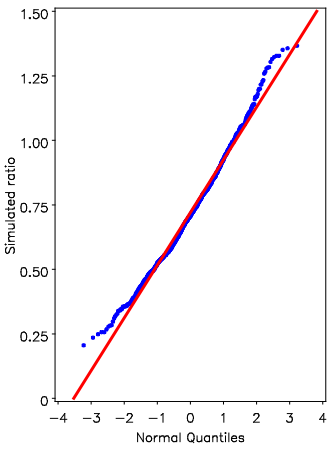
$n_j = 50$

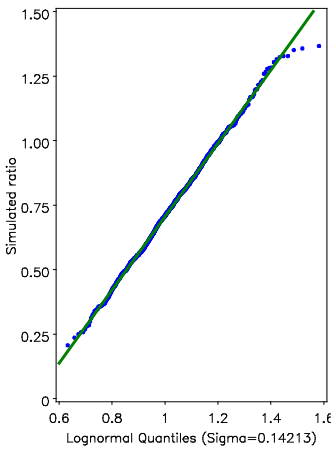


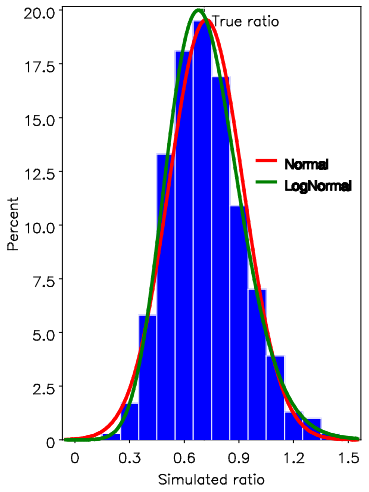




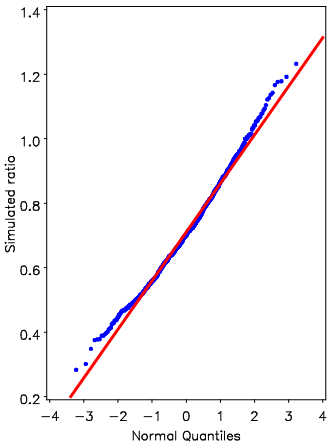
$n_j = 100$

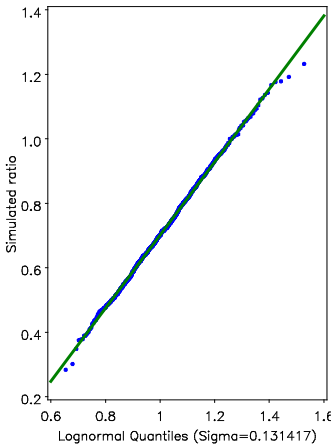






$n_j = 200$





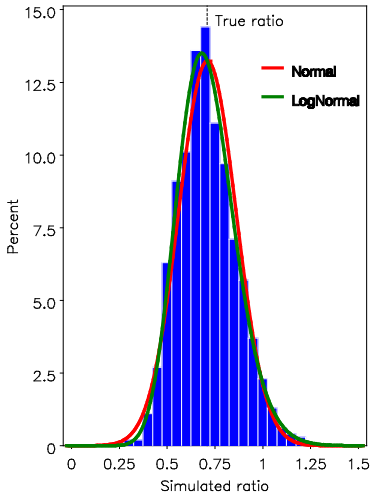


Figure F.2 Upper limits of SMR 95% confidence (credible) intervals by method: SMR 1.37, n_j = 100 and n_R = 1000

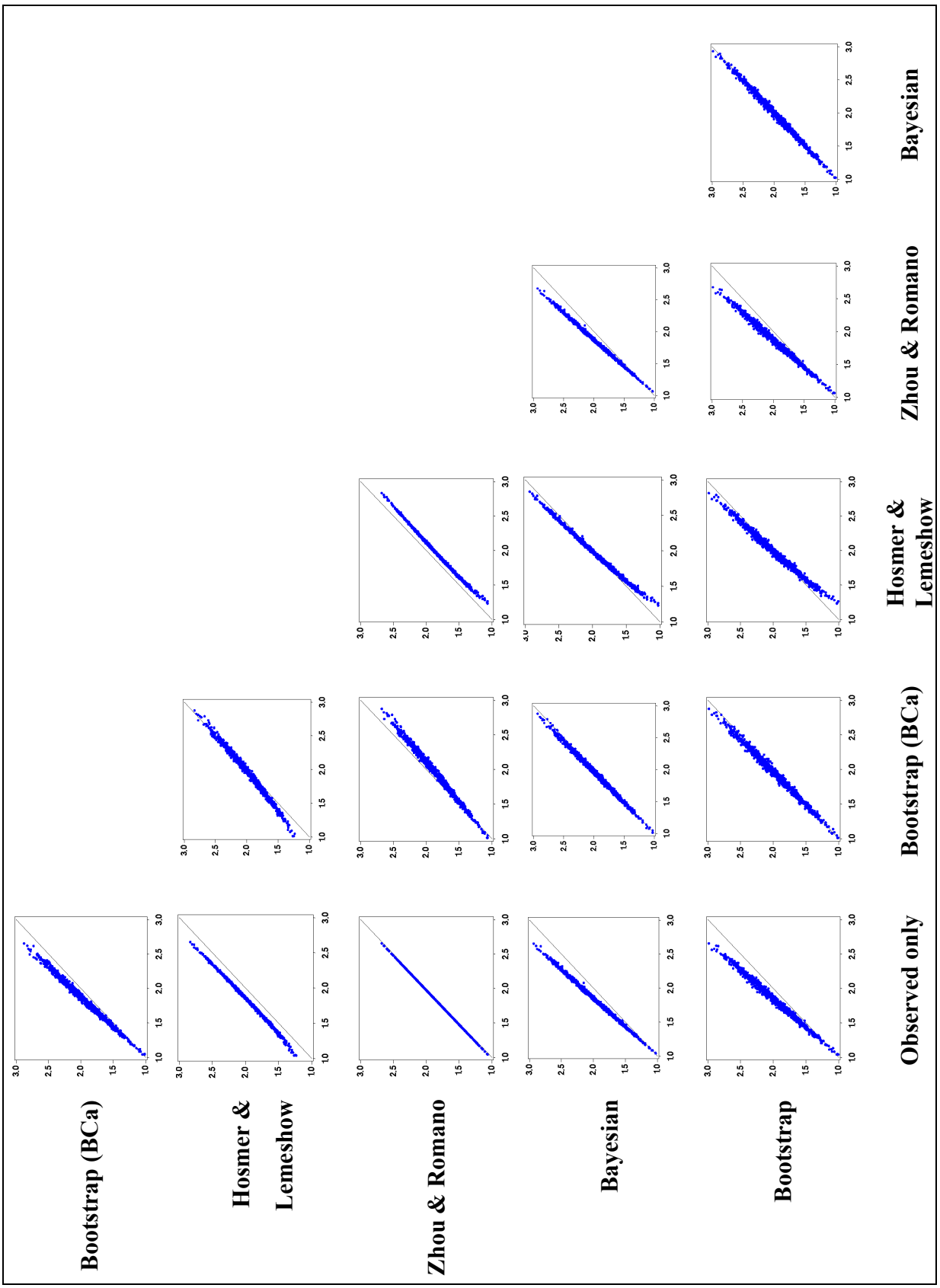


Figure F.3 Upper limits of SMR 95% confidence intervals by method: SMR ≈ 0.71, n_j = 100 and n_R = 1000

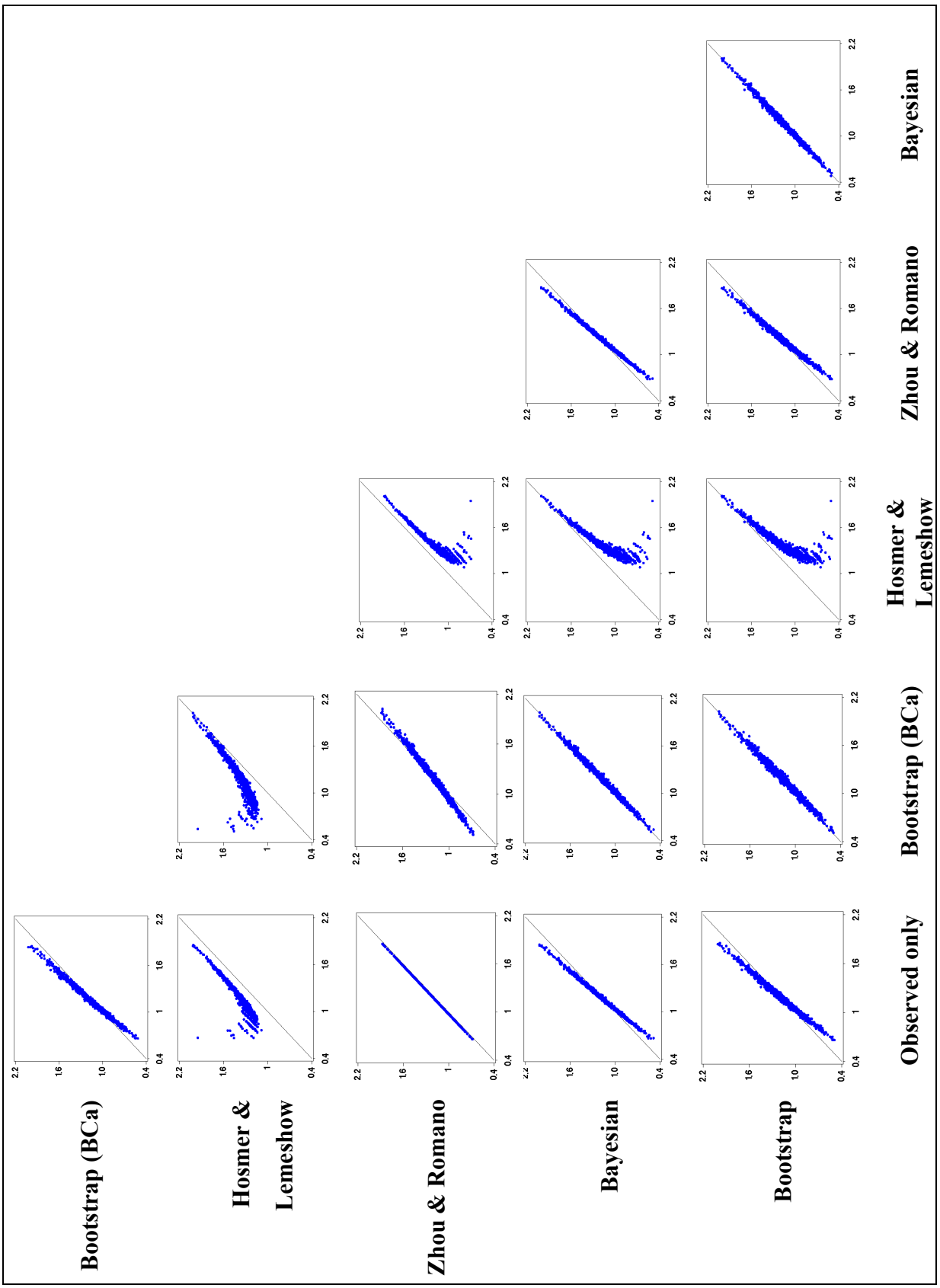


Figure F.4 Lower limits of SMR 95% confidence intervals by method: SMR ≈ 1.37, n_j = 100 and n_R = 1000

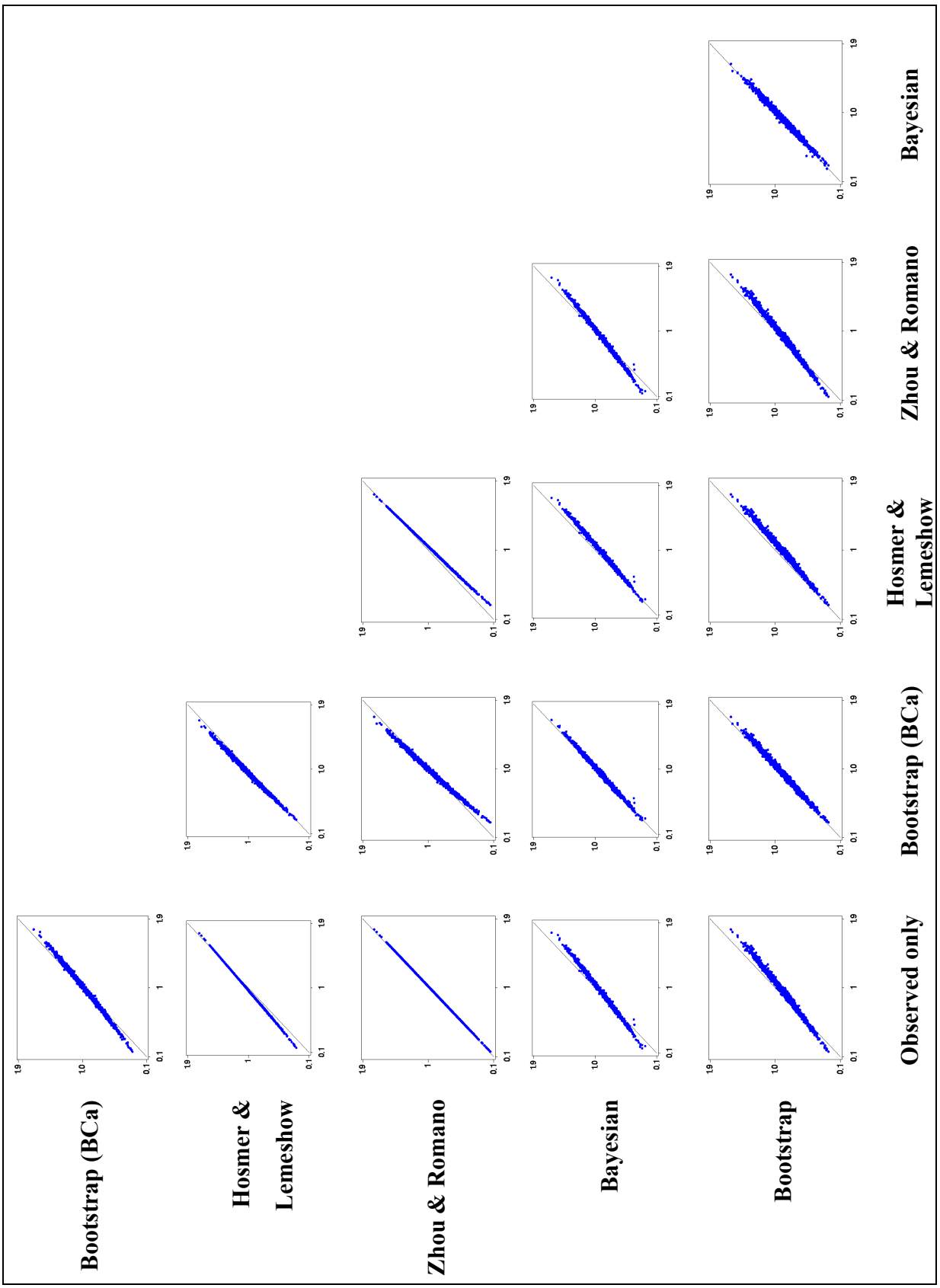
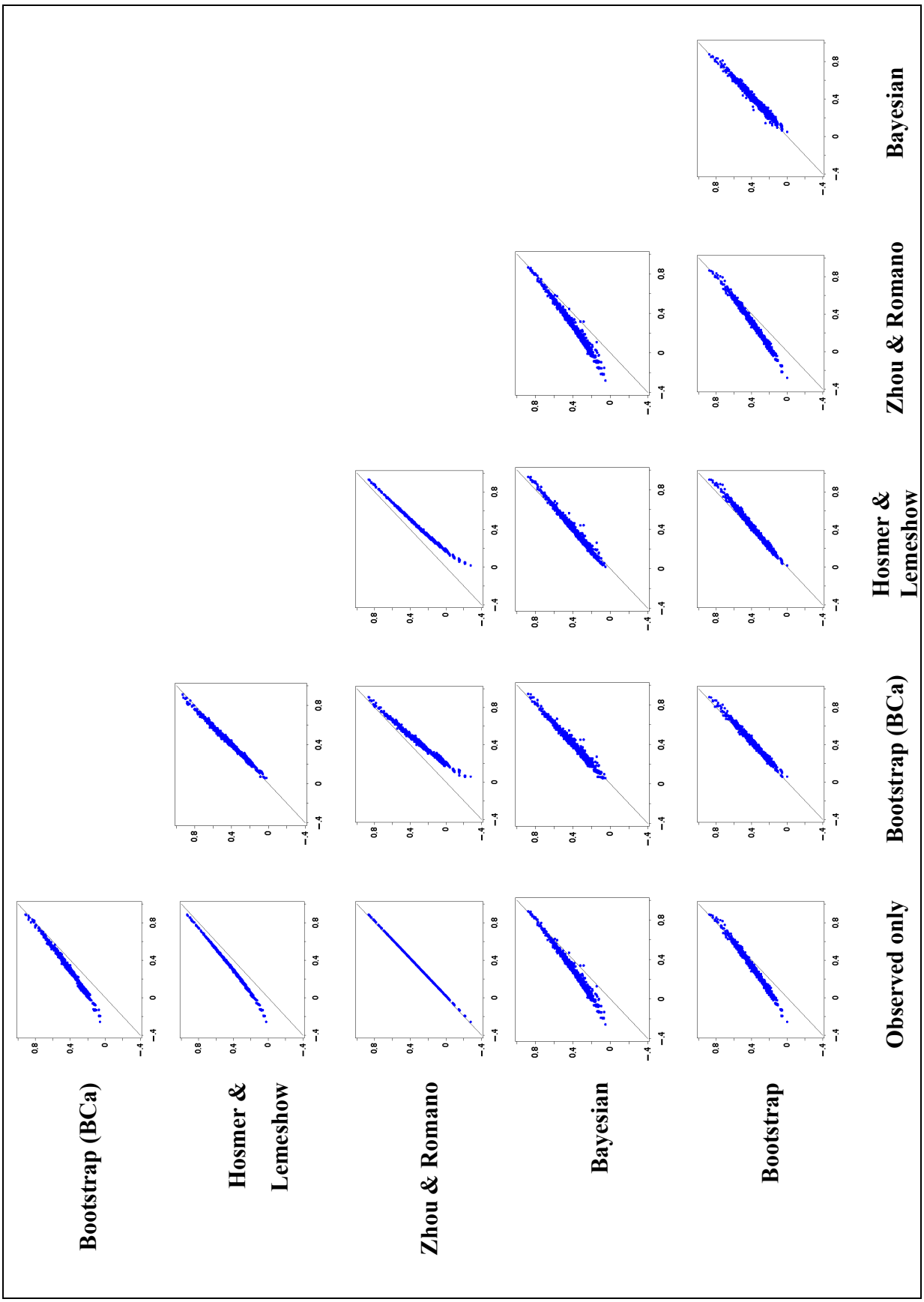


Figure F.5 Lower limits of SMR 95% confidence intervals by method: SMR ≈ 0.71, n_j = 100 and n_R = 1000



Appendix F.3 Coverage of estimated intervals

Tables VI.ii to VI.iv show the proportion of simulated intervals that did not contain the value of the true SMR for each scenario.

Table F.2 Coverage of estimated 95% confidence interval: SMR = 1

n_j	Source of error	Method	n_R 500	1000	2000
50	Observed only	Normal (with CC)	0.042	0.037	0.027
		Normal (without CC)	0.066	0.063	0.047
		Normal ('full')	0.038	0.032	0.024
		BCa bootstrap	0.075	0.067	0.059
	Observed & Expected	Hosmer & Lemeshow	0.067	0.065	0.051
		Zhou & Romano	0.057	0.057	0.047
		Bayesian	0.074	0.070	0.061
		Bootstrap	0.066	0.063	0.066
100	Observed only	Normal (with CC)	0.046	0.050	0.034
		Normal (without CC)	0.065	0.067	0.047
		Normal ('full')	0.040	0.044	0.030
		BCa bootstrap	0.061	0.067	0.047
	Observed & Expected	Hosmer & Lemeshow	0.047	0.051	0.047
		Zhou & Romano	0.037	0.055	0.042
		Bayesian	0.043	0.065	0.048
		Bootstrap	0.042	0.063	0.047
200	Observed only	Normal (with CC)	0.080	0.045	0.048
		Normal (without CC)	0.093	0.059	0.060
		Normal ('full')	0.072	0.038	0.042
		BCa bootstrap	0.086	0.059	0.056
	Observed & Expected	Hosmer & Lemeshow	0.042	0.042	0.046
		Zhou & Romano	0.049	0.034	0.048
		Bayesian	0.045	0.036	0.054
		Bootstrap	0.047	0.036	0.052

Table F.3 Coverage of estimated 95% confidence interval: SMR ≈ 1.37

n_j	Source of error	Method	n_R 500	1000	2000
50	Observed only	Normal (with CC)	0.090	0.059	0.056
		Normal (without CC)	0.128	0.093	0.092
		Normal ('full')	0.053	0.034	0.035
		BCa bootstrap	0.089	0.064	0.062
	Observed & Expected	Hosmer & Lemeshow	0.073	0.060	0.085
		Zhou & Romano	0.108	0.082	0.086
		Bayesian	0.088	0.068	0.066
		Bootstrap	0.074	0.065	0.056
100	Observed only	Normal (with CC)	0.087	0.050	0.070
		Normal (without CC)	0.113	0.129	0.096
		Normal ('full')	0.055	0.055	0.044
		BCa bootstrap	0.076	0.090	0.057
	Observed & Expected	Hosmer & Lemeshow	0.055	0.078	0.068
		Zhou & Romano	0.080	0.114	0.083
		Bayesian	0.058	0.073	0.056
		Bootstrap	0.052	0.066	0.055
200	Observed only	Normal (with CC)	0.171	0.105	0.081
		Normal (without CC)	0.190	0.122	0.102
		Normal ('full')	0.113	0.071	0.046
		BCa bootstrap	0.129	0.085	0.058
	Observed & Expected	Hosmer & Lemeshow	0.072	0.073	0.066
		Zhou & Romano	0.120	0.096	0.081
		Bayesian	0.061	0.054	0.044
		Bootstrap	0.061	0.051	0.045

Table F.4 Coverage of estimated 95% confidence interval: SMR ≈ 0.71

n_j	Source of error	Method	n_R 500	1000	2000
50	Observed only	Normal (with CC)	0.017	0.012	0.007
		Normal (without CC)	0.025	0.018	0.012
		Normal ('full')	0.026	0.019	0.014
		BCa bootstrap	0.059	0.050	0.051
	Observed & Expected	Hosmer & Lemeshow	0.046	0.039	0.031
		Zhou & Romano	0.018	0.017	0.011
		Bayesian	0.076	0.066	0.082
		Bootstrap	0.072	0.068	0.079
100	Observed only	Normal (with CC)	0.024	0.022	0.028
		Normal (without CC)	0.034	0.030	0.034
		Normal ('full')	0.045	0.035	0.043
		BCa bootstrap	0.068	0.057	0.070
	Observed & Expected	Hosmer & Lemeshow	0.040	0.040	0.038
		Zhou & Romano	0.022	0.026	0.034
		Bayesian	0.060	0.060	0.068
		Bootstrap	0.057	0.063	0.065
200	Observed only	Normal (with CC)	0.028	0.028	0.021
		Normal (without CC)	0.041	0.033	0.035
		Normal ('full')	0.052	0.047	0.047
		BCa bootstrap	0.068	0.065	0.059
	Observed & Expected	Hosmer & Lemeshow	0.025	0.031	0.028
		Zhou & Romano	0.017	0.020	0.022
		Bayesian	0.047	0.053	0.059
		Bootstrap	0.040	0.053	0.055

More details on the coverage properties of the methods are given in Tables F.5 to F.12. The proportion of intervals falling wholly above and below the true value for the SMR are reported for each method.

Table F.5 Normal Approximation (with continuity correction)

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.010	0.027	0.007
	Over	0.032	0.010	0.020
100	Under	0.011	0.018	0.013
	Over	0.035	0.032	0.021
200	Under	0.031	0.018	0.021
	Over	0.049	0.027	0.027
Ratio ≈ 1.37				
50	Under	0.037	0.026	0.026
	Over	0.053	0.033	0.040
100	Under	0.036	0.050	0.030
	Over	0.051	0.054	0.040
200	Under	0.065	0.047	0.033
	Over	0.106	0.058	0.048
Ratio ≈ 0.71				
50	Under	0.000	0.000	0.000
	Over	0.017	0.012	0.007
100	Under	0.008	0.004	0.011
	Over	0.024	0.018	0.017
200	Under	0.009	0.010	0.010
	Over	0.019	0.018	0.011

Table F.6 Normal Approximation (without continuity correction)

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.019	0.022	0.018
	Over	0.047	0.041	0.029
100	Under	0.019	0.026	0.020
	Over	0.046	0.041	0.027
200	Under	0.035	0.021	0.028
	Over	0.058	0.038	0.032
Ratio ≈ 1.37				
50	Under	0.059	0.046	0.039
	Over	0.069	0.047	0.053
100	Under	0.052	0.064	0.046
	Over	0.061	0.065	0.050
200	Under	0.078	0.054	0.039
	Over	0.112	0.068	0.063
Ratio ≈ 0.71				
50	Under	0.003	0.001	0.001
	Over	0.022	0.017	0.011
100	Under	0.010	0.007	0.015
	Over	0.024	0.023	0.019
200	Under	0.016	0.013	0.016
	Over	0.025	0.020	0.019

Table F.7 Normal Approximation ("full")

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.008	0.007	0.006
	Over	0.030	0.025	0.018
100	Under	0.010	0.015	0.011
	Over	0.030	0.029	0.019
200	Under	0.028	0.014	0.017
	Over	0.044	0.024	0.025
Ratio ≈ 1.37				
50	Under	0.015	0.011	0.015
	Over	0.038	0.023	0.020
100	Under	0.022	0.021	0.018
	Over	0.033	0.034	0.026
200	Under	0.043	0.031	0.014
	Over	0.070	0.040	0.032
Ratio ≈ 0.71				
50	Under	0.003	0.001	0.002
	Over	0.023	0.018	0.012
100	Under	0.014	0.008	0.015
	Over	0.031	0.027	0.028
200	Under	0.020	0.017	0.021
	Over	0.032	0.030	0.026

Table F.8 BCa Bootstrap

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.034	0.034	0.035
	Over	0.041	0.033	0.024
100	Under	0.024	0.032	0.027
	Over	0.037	0.035	0.020
200	Under	0.036	0.028	0.030
	Over	0.050	0.031	0.026
Ratio ≈ 1.37				
50	Under	0.042	0.035	0.031
	Over	0.048	0.029	0.031
100	Under	0.034	0.048	0.030
	Over	0.042	0.042	0.027
200	Under	0.051	0.039	0.024
	Over	0.078	0.046	0.034
Ratio ≈ 0.71				
50	Under	0.029	0.026	0.037
	Over	0.030	0.024	0.014
100	Under	0.038	0.029	0.037
	Over	0.030	0.028	0.033
200	Under	0.036	0.027	0.031
	Over	0.032	0.038	0.028

Table F.9 Hosmer & Lemeshow

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.000	0.000	0.000
	Over	0.067	0.065	0.051
100	Under	0.000	0.000	0.000
	Over	0.047	0.045	0.047
200	Under	0.006	0.006	0.006
	Over	0.036	0.036	0.040
Ratio ≈ 1.37				
50	Under	0.000	0.000	0.000
	Over	0.073	0.060	0.085
100	Under	0.009	0.013	0.006
	Over	0.046	0.065	0.062
200	Under	0.020	0.022	0.011
	Over	0.052	0.051	0.055
Ratio ≈ 0.71				
50	Under	0.000	0.000	0.000
	Over	0.046	0.039	0.031
100	Under	0.000	0.000	0.000
	Over	0.040	0.040	0.038
200	Under	0.000	0.000	0.000
	Over	0.025	0.031	0.028

Table F.10 Zhou & Romano

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.016	0.021	0.018
	Over	0.041	0.036	0.029
100	Under	0.009	0.021	0.018
	Over	0.028	0.027	0.024
200	Under	0.017	0.013	0.021
	Over	0.032	0.021	0.027
Ratio ≈ 1.37				
50	Under	0.047	0.040	0.035
	Over	0.061	0.042	0.051
100	Under	0.033	0.055	0.039
	Over	0.047	0.059	0.044
200	Under	0.044	0.041	0.033
	Over	0.076	0.055	0.048
Ratio ≈ 0.71				
50	Under	0.000	0.001	0.001
	Over	0.018	0.016	0.010
100	Under	0.007	0.005	0.015
	Over	0.015	0.021	0.019
200	Under	0.003	0.006	0.010
	Over	0.014	0.014	0.012

Table F.11 Bayesian

n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.044	0.042	0.041
	Over	0.030	0.028	0.020
100	Under	0.023	0.036	0.030
	Over	0.020	0.029	0.018
200	Under	0.026	0.020	0.032
	Over	0.019	0.016	0.022
Ratio ≈ 1.37				
50	Under	0.050	0.035	0.031
	Over	0.038	0.029	0.031
100	Under	0.029	0.043	0.032
	Over	0.029	0.030	0.024
200	Under	0.025	0.030	0.019
	Over	0.036	0.024	0.025
Ratio ≈ 0.71				
50	Under	0.052	0.040	0.064
	Over	0.024	0.026	0.021
100	Under	0.038	0.036	0.049
	Over	0.022	0.024	0.027
200	Under	0.027	0.030	0.036
	Over	0.020	0.023	0.023

Table F.12 Bootstrap

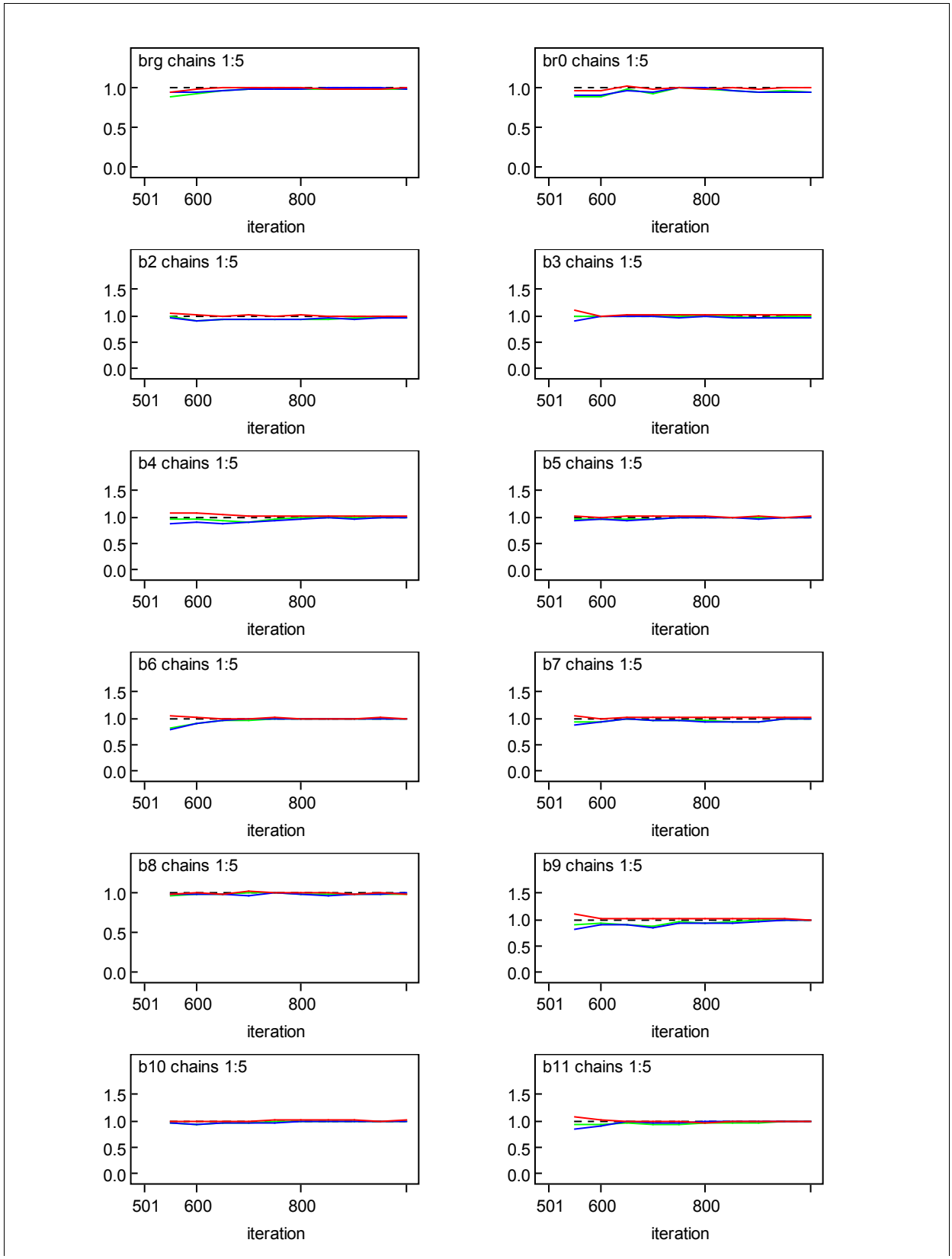
n_j		n_R 500	1000	2000
Ratio = 1				
50	Under	0.047	0.044	0.048
	Over	0.019	0.019	0.018
100	Under	0.026	0.037	0.032
	Over	0.016	0.026	0.015
200	Under	0.027	0.023	0.034
	Over	0.020	0.013	0.018
Ratio ≈ 1.37				
50	Under	0.048	0.044	0.038
	Over	0.026	0.021	0.018
100	Under	0.026	0.045	0.035
	Over	0.026	0.021	0.020
200	Under	0.029	0.030	0.026
	Over	0.032	0.021	0.019
Ratio ≈ 0.71				
50	Under	0.058	0.057	0.072
	Over	0.014	0.011	0.007
100	Under	0.042	0.043	0.041
	Over	0.015	0.020	0.024
200	Under	0.027	0.032	0.039
	Over	0.013	0.021	0.016

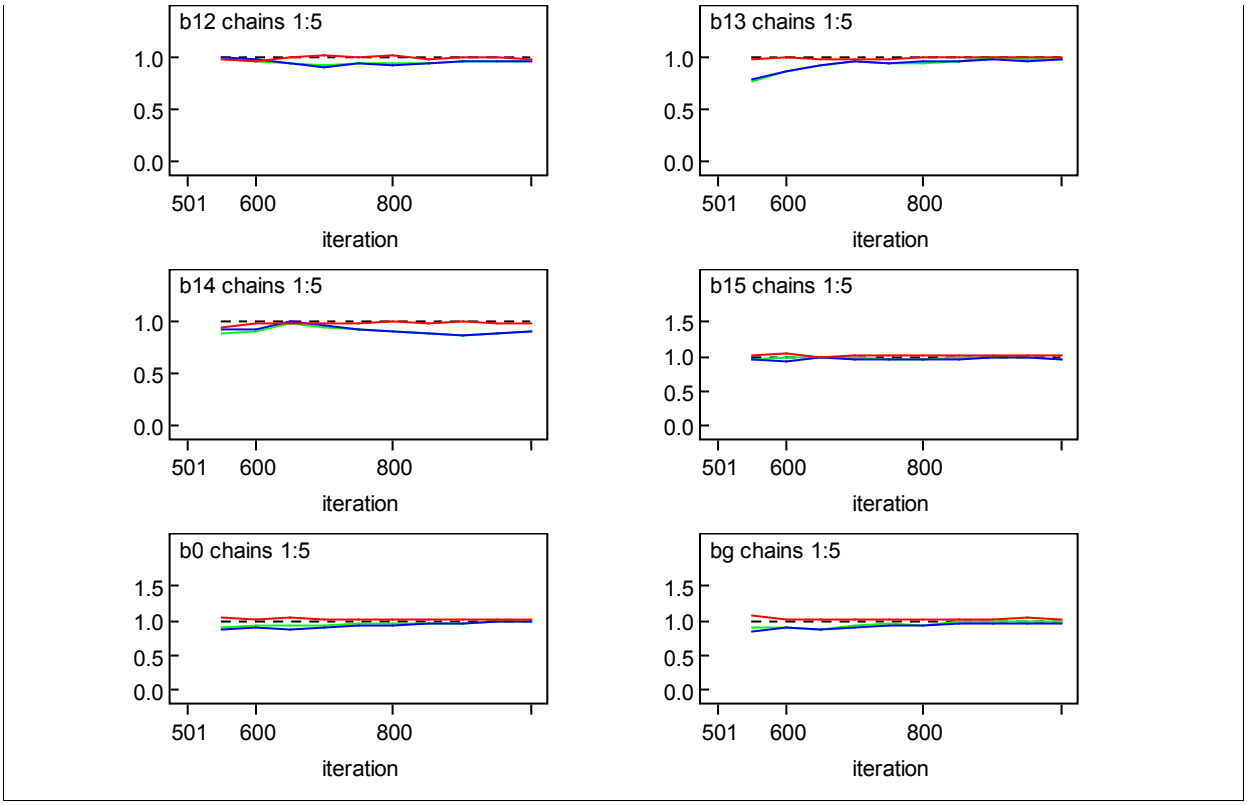
Appendix F.4 Further details of Bayesian estimate of SMR for TNS data

This is an example of the models used to investigate the sensitivity of the values of the model parameters and the SMR to the choice of prior probability distributions in §5.8.

```
model ci {  for (i in 101:3025) {          # Reference data
            died[i] ~ dbern(p[i])
            c_gest[i] <- gest[i]-30        # Centre gestational age

      # Estimate model parameters from reference data
            logit(p[i]) <-    br0  +  b2*i2[i]  +  b3*i3[i]  +  b4*i4[i]  +
                        b5*i5[i]  +  b6*i6[i]  +  b7*i7[i]  +  b8*i8[i]  +
                        b9*i9[i] + b10*i10[i] + b11*i11[i] + b12*i12[i] +
                        b13*i13[i]    +    b14*i14[i]    +    b15*i15[i]    +
                        brg*c_gest[i]
 }

for (i in 1:100) {                         # Data from unit of interest

            c_gest[i] <- gest[i]-30        # Centre gestational age

      # Calculate expected 'p' using br0 & brg estimated above
            logit(pp[i]) <- br0 + brg*c_gest[i]

      # Estimate model parameters b0 & bg from unit of interest
            died[i] ~ dbern(op[i])
            logit(op[i]) <- b0 + bg*c_gest[i]
 }
# Calculate SMR
      sum.pp <- sum(pp[])                  # Sum of predicted
      sum.ob <- sum(op[])                  # Sum of observed
      ratio <- sum.ob/sum.pp               # Ratio of interest

# Prior distributions
      b0 <- logit(g)
      br0 <- logit(gr)
      g ~ dbeta(0.25,4.75)
      gr ~ dbeta(0.25,4.75)

      brg ~ dnorm(0,1.0E-6)
      bg ~ dnorm(0,1.0E-6)

      b2 ~ dnorm(0,1.0E-6)
      b3 ~ dnorm(0,1.0E-6)
      b4 ~ dnorm(0,1.0E-6)
      b5 ~ dnorm(0,1.0E-6)
      b6 ~ dnorm(0,1.0E-6)
      b7 ~ dnorm(0,1.0E-6)
      b8 ~ dnorm(0,1.0E-6)
      b9 ~ dnorm(0,1.0E-6)
      b10 ~ dnorm(0,1.0E-6)
      b11 ~ dnorm(0,1.0E-6)
      b12 ~ dnorm(0,1.0E-6)
      b13 ~ dnorm(0,1.0E-6)
      b14 ~ dnorm(0,1.0E-6)
      b15 ~ dnorm(0,1.0E-6) }
```

Figure F.6 Plots of Brooks-Gelman-Rubin statistic for Unit 16





There was a potential problem in the estimation of β_0 and β_G for Units 3 and 9 because of quasi-complete separation of the data. Trace plots for Unit 3 are reproduced, from the third scenario, below showing that within the 1,000 iteration burn-in the five chains were all sampling from the same sample space well before the end of the burn-in (Figure F.7). This was confirmed by plots of the Brooks-Gelman-Rubin statistic (Figure E.4). Similar plots were obtained for Unit 9.

Figure F.7 Trace plots for burn-in: Unit 3

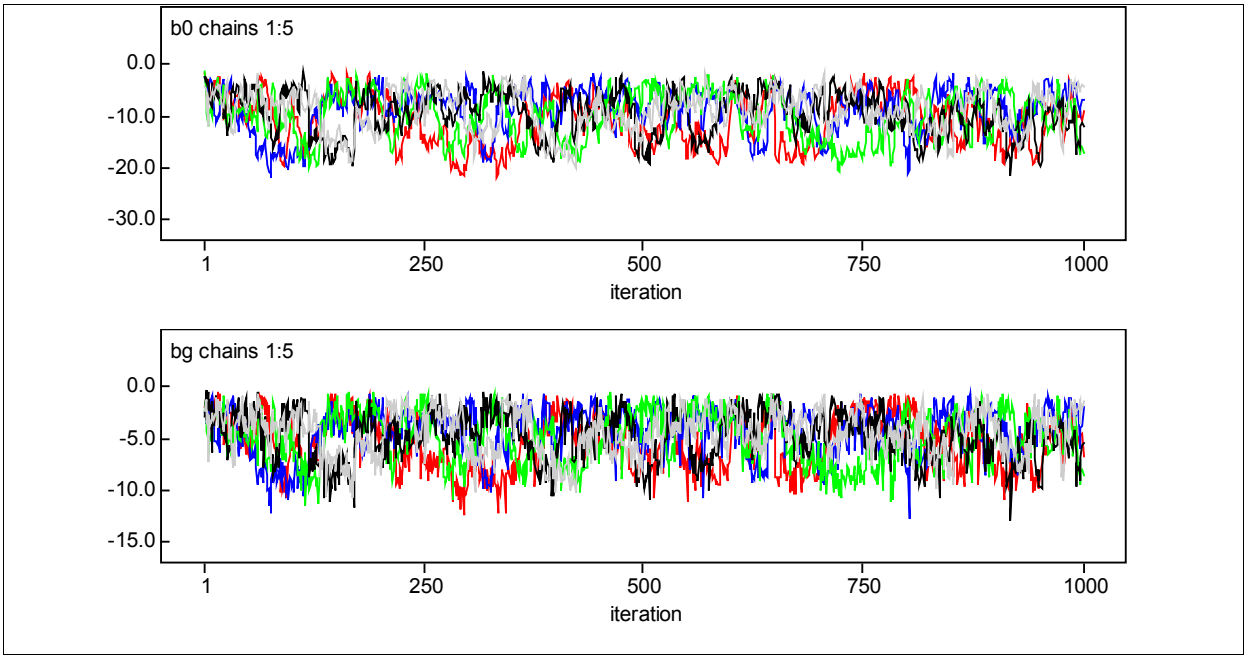
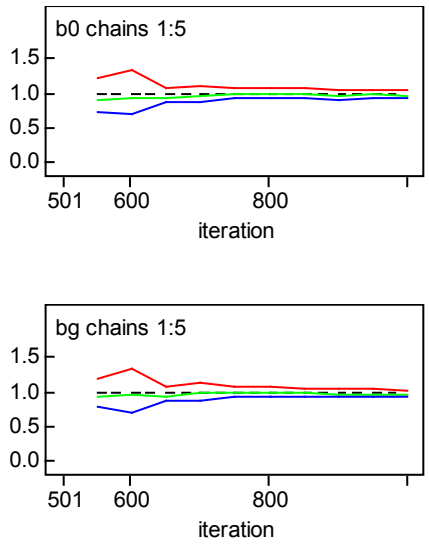


Figure F.8 Brooks-Gelman-Rubin statistic plots for burn-in: Unit 3

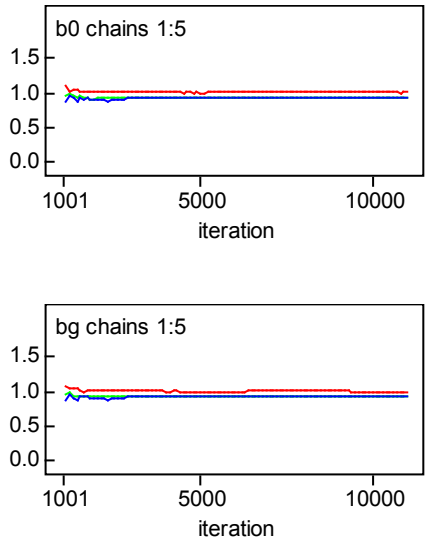


Last bin: $R = 1.050$

Last bin: $R = 1.034$

The five chains were then run for a further 10,000 iterations and the Brooks-Gelman-Rubin statistics calculated. Once again there was no evidence that the chains were no sampling from the same distribution.

Figure F.9 Brooks-Gelman-Rubin statistic plots for sampled iteration: Unit 3



First bin: $R = 1.089$

Last bin: $R = 1.002$

First bin: $R = 1.075$

Last bin: $R = 1.001$

Table F.13 Parameter estimates using various prior distributions (Unit 3)

Prior distibution					**Parameter Estimate**							
β_{R0}	β_{RG}	β_K	β_0	β_G	$\hat{\beta}_{R0}$	$\hat{\beta}_{RG}$	$\hat{\beta}_0$	$\hat{\beta}_G$	$\hat{\Sigma d_i}$	$\hat{\Sigma \pi_i}$	$\hat{\text{SMR}}$	**(95% CI)**
Frequentist model					-3.73	-0.66	-18.57	-9.29	2	1.11	1.80	
$\pi \sim U(0,1)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim U(0,1)$	$N(0, 1000^2)$	-4.05	-0.65	-3.37	-1.66	2.79	1.07	2.61	(1.15 to 5.14)
$\pi \sim B(0.25, 4.75)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim B(0.25, 4.75)$	$N(0, 1000^2)$	-4.09	-0.59	-6.66	-3.30	2.17	1.06	2.04	(0.98 to 3.68)
$\pi \sim B(0.1, 1.9)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim B(0.1, 1.9)$	$N(0, 1000^2)$	-4.09	-0.66	-9.58	-4.82	2.07	1.05	1.96	(0.97 to 3.40)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.01	-0.66	-9.49	-4.76	2.07	1.06	1.98	(0.99 to 3.44)
$N(-2.94, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(-2.94, 1000^2)$	$N(0, 1000^2)$	-4.09	-0.66	-10.68	-5.35	2.05	1.05	1.95	(0.98 to 3.33)
$N(-2.94, 0.80)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(-2.94, 0.80)$	$N(0, 1000^2)$	-4.06	-0.65	-3.72	-1.77	2.55	1.06	2.42	(1.18 to 4.29)
$U(-25, 25)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$U(-25, 25)$	$N(0, 1000^2)$	-4.09	-0.66	-14.72	-7.35	2.03	1.05	1.93	(0.96 to 3.24)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(\beta_{R0}, 1000^2)$	$N(\beta_{RG}, 1000^2)$	-4.09	-0.66	-11.54	-5.77	2.06	1.06	1.95	(0.97 to 3.32)
$N(0, 1000^2)$	$N(0, 1000^2)$	$U(-2, 2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.05	-0.65	-10.05	-5.07	2.09	1.07	1.94	(0.97 to 3.31)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.00	-0.65	-9.89	-4.94	2.06	1.11	1.85	(0.94 to 3.14)

Table F.14 Parameter estimates using various prior distributions (Unit 8)

Prior distibution					**Parameter Estimate**							
β_{R0}	β_{RG}	β_K	β_0	β_G	$\hat{\beta}_{R0}$	$\hat{\beta}_{RG}$	$\hat{\beta}_0$	$\hat{\beta}_G$	$\Sigma\hat{d}_i$	$\Sigma\hat{\pi}_i$	$\hat{\text{SMR}}$	(95% CI)
Frequentist model					-3.73	-0.67	-3.92	-0.58	8	9.8	0.82	
$\pi \sim U(0,1)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim U(0,1)$	$N(0, 1000^2)$	-4.01	-0.66	-3.64	-0.52	8.76	9.31	0.94	(0.49 to 1.63)
$\pi \sim B(0.25, 4.75)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim B(0.25, 4.75)$	$N(0, 1000^2)$	-4.05	-0.66	-4.03	-0.60	7.97	9.13	0.87	(0.42 to 1.53)
$\pi \sim B(0.1, 1.9)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$\pi \sim B(0.1, 1.9)$	$N(0, 1000^2)$	-4.04	-0.66	-4.07	-0.61	7.85	9.17	0.86	(0.41 to 1.53)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.05	-0.66	-4.15	-0.62	7.79	9.15	0.86	(0.42 to 1.51)
$N(-2.94, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(-2.94, 1000^2)$	$N(0, 1000^2)$	-4.05	-0.66	-4.12	-0.61	7.86	9.17	0.86	(0.41 to 1.52)
$N(-2.94, 0.80)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(-2.94, 0.80)$	$N(0, 1000^2)$	-4.02	-0.66	-3.74	-0.54	8.49	9.25	0.92	(0.47 to 1.57)
$U(-25, 25)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$U(-25, 25)$	$N(0, 1000^2)$	-4.04	-0.66	-4.11	-0.62	7.84	9.16	0.86	(0.42 to 1.53)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	$N(\beta_{R0}, 1000^2)$	$N(\beta_{RG}, 1000^2)$	-4.05	-0.66	-4.12	-0.62	7.84	9.17	0.86	(0.41 to 1.51)
$N(0, 1000^2)$	$N(0, 1000^2)$	$U(-2, 2)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-4.01	-0.66	-4.10	-0.61	7.84	9.32	0.84	(0.41 to 1.47)
$N(0, 1000^2)$	$N(0, 1000^2)$	$N(0, 1)$	$N(0, 1000^2)$	$N(0, 1000^2)$	-3.97	-0.66	-4.09	-0.61	7.89	9.59	0.82	(0.40 to 1.43)

Appendix G: RISK-ADJUSTMENT VARIABLES

This appendix gives additional details of the investigation into potential risk-adjustment variables summarised in §6.4.

Maximum likelihood estimates of the model parameters, and odds ratios where there is no variable interaction, are reported. The heterogeneity of the effect across NICUs was tested by adding an indicator variable into the logistic model representing NICU of care and an interaction between NICU and the variable of interest. This is reported using the p-value for the interaction term. In addition, SMRs for each unit are reported using the deviation parameterization models described in Chapter 5 and 95% confidence intervals estimated using the method of Hosmer & Lemeshow (1995) (§5.6.2).

The area under the ROC curve (A_{ROC}) and the Hosmer & Lemeshow goodness-of-fit test statistic ($\hat{C}$), and associated p-value, are reported for each model, but more detailed model checking was considered in §6.6 where a final model was developed.

Appendix G.1 Gestational age at birth

For preterm births the gestational age of an infant at birth has a strong monotonic relationship with neonatal mortality (Verloove-Vanhorick *et al,* 1986). However, in almost all cases the exact day of conception is unknown and, therefore, must be estimated. There are three approaches to estimating the gestational age of an infant: mother's dates, ultrasound scan and postnatal examination. Each of these is discussed below.

Mothers' dates

Although the standard definition of gestational age is the time from the onset of the last normal menses to the date of birth, the estimation of gestational age is not straightforward. The reported date of onset of the Last Menstrual Period (LMP) before pregnancy can be incorrect for a number of reasons, including irregularities of the menstrual cycle, individual variations in the length of the cycle, preconception amenorrhea following oral contraceptives, implantation bleeding or other bleeding in pregnancy, and recall error by the mother (Gjessing *et al,* 1999). Such discrepancies are more common among preterm (and post-term) births than with term births (Kramer *et al,* 1988; Mustafa and David, 2001). In addition, there is some evidence for digit preference in reported LMP dates amongst women (Waller *et al,* 2000). In

particular, the 15th was recorded as the date of onset of LMP two-and-a-half times more than expected, but such number preferences are likely to introduce only small discrepancies into the data. There is also evidence that errors in gestational age estimated using the mothers' dates are more frequent in certain groups; e.g. younger mothers and smokers (Savitz *et al,* 2002; Yang *et al,* 2002).

Ultrasound scan

To try to overcome such problems in using the date of onset of the last menstrual period, early (≤ 20 weeks) ultrasound measurements of fetal dimensions can be used to estimate gestational age. Such scans are now part of routine antenatal care in the United Kingdom and there is evidence that the routine use of scans has clinical benefits (Neilson, 2003). To estimate gestational age, it is recommended that fetal crown-rump length be used at 10-13 weeks and the biparietal diameter (BPD) used beyond 14 weeks gestational age (National Collaborating Centre for Women's and Children's Heath, 2003:34).

However, although such measurements are useful, they, too, are only estimates of true gestational age. Their accuracy is dependent on many factors, including the charts used to estimate gestational age from fetal measurements (Altman and Chitty, 1997; Hadlock *et al,* 1982; Eriksen *et al,* 1985; Campbell and Newman, 1971), the accuracy of the measurements taken by the radiographer and even the position of the fetus (Altman and Chitty, 1997). The assumption is also made that all fetuses of the same size are of the same gestational age and, therefore, any variation in fetal size will be interpreted as differences in gestational age (Henriksen *et al,* 1995). Although BPD has a smaller inter-individual variation than other fetal measurements, there is still some variation between fetuses of the same (true) gestational age (Rabelink *et al,* 1994). Female fetuses tend to be smaller than male at the same age and so are more likely to be judged younger (e.g. by an average 2.5 days (Kallen, 1995) or 1.5 days (Tunon *et al,* 1999)), as are fetuses of younger mothers, smokers, multiparous women and women with low educational levels (Källén, 2002). The tables by Altman and Chitty (1997) report an uncertainly in the estimation of gestational age using BPD of some 10 to 12 days for fetuses estimated to be of 16 to 19 weeks gestational age.

As well as such random errors, there is also evidence for a systematic difference between gestational ages estimated using LMP dates compared to those estimated from fetal measurements. An amenorrheic cycle just before conception, or spotting around the expected time of the first missed period, can result in estimates of gestational age that are either one

month too large or too small (Joseph *et al,* 2001). However, some studies have found only a small proportion of such discrepancies (Altman and Chitty, 1997; Yang *et al,* 2002). The more general trend is that, even for women with a normal last menstrual period, estimates of gestational age from ultrasound scans tend to be lower that those from LMP dates (Oja *et al,* 1991; Mustafa and David, 2001; Yang *et al,* 2002). Such discrepancies are, on average, quite small, e.g. 2.8 days (Savitz *et al,* 2002), and are probably due to delayed ovulation (> 14 days), which is more common that early ovulation (< 14 days) (Yang *et al,* 2002).

Although various methods have been derived to try to combine information from both LMP dates and BPD estimates (Blondel *et al,* 2002), these have not been used with the data in the Trent Neonatal Survey.

Postnatal examination

Before estimates of gestational age were routinely available from ultrasound scans, and currently when such scans have not been performed, and where there was uncertainty over the LMP dates, gestational age at birth was estimated by physical examination of the infant. Sola and Chow have published an interesting short review of the history of such methods (Sola and Chow, 1999). Of the methods proposed two predominate: the Dubowitz Scale (Dubowitz *et al,* 1970) and the New Ballard Score (Ballard *et al,* 1991), an expansion of the original Ballard Maturation Score (Ballard *et al,* 1979) to include preterm infants. One problem with such methods is that they assume that all fetuses develop at the same rate. It has been shown that these scores perform poorly with preterm infants, with gestational age estimated using these scores generally exceeded that estimated using known LMP dates: mean differences ranging from 1.3 to 3.3 weeks (Donovan *et al,* 1999; Wariyar *et al,* 1997).

Estimation of gestational age in TNS data

The estimates of gestational age recorded by the Trent Neonatal Survey are those taken from the mothers' and infants' notes. The estimate used, therefore, depends on the policy of each obstetric unit. However the general procedure followed within the Region is to use the following hierarchy (Bohin *et al,* 1999):

v) Mother certain of her dates (most reliable);

vi) Early dating scan

vii) Late dating scan;

viii) Postnatal examination (least reliable).

This aims to use the 'best available' estimates. Previous work with the Trent Neonatal Survey data has shown that about 36% of the gestational ages were amended, usually to the estimate from an early scan, before being recorded in the survey (Draper *et al,* 1999). The proportion amended in the data used for this thesis is unknown.

In addition, there is the potential for differential measurement error between the units, either because different dating methods are used or because of the different interpretation of ultrasound measurements. The extent of such errors is not known, but since reported errors between gestational age estimated from mothers' dates and scan data are only in the order of a few days (described above) this is unlikely to be the explanation of any differences found between the units.

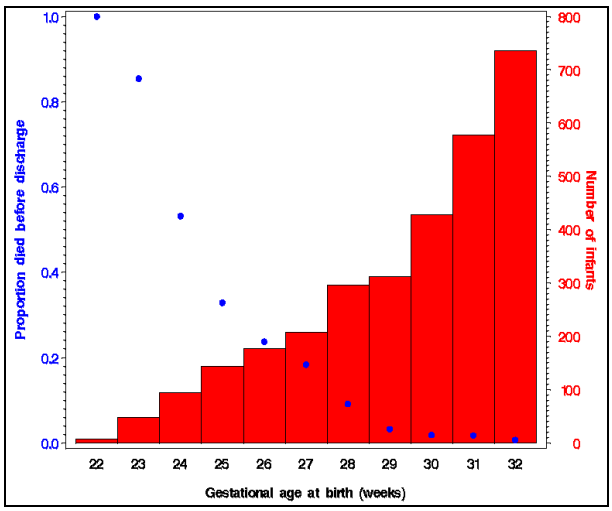
Association with in-unit mortality

The proportions of infants dying before discharge by recorded gestational age at birth are shown in Table G.1. Perhaps this is more clearly illustrated in Figure G.1. Unsurprisingly, there is a clear trend for lower mortality with increasing gestational age.

Table G.1 Observed mortality by gestational age at birth

Gestational age	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
No. infants	7	48	94	143	177	207	296	312	428	578	735
No. died	7	41	50	47	42	38	27	10	8	10	5
Proportion	1.00	0.85	0.53	0.33	0.24	0.18	0.09	0.03	0.02	0.02	0.01

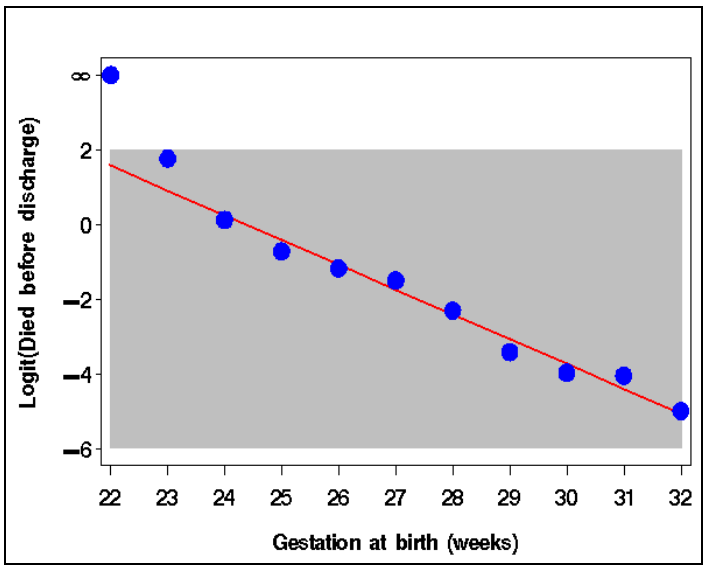
Figure G.1 Observed Mortality by Gestational Age at Birth



When modelled within a logistic regression model, there is strong statistical evidence for such a relationship. Although comprehensive model checking will not be discussed until §6.6, an inspection of the estimated function shows an approximately linear relationship between the logit of the probability of death before discharge and gestational age at birth. There was a suggestion that at very early gestations (22 and 23 weeks) the linear function underestimates the true mortality. However, the addition of higher value polynomials did not produce a statistically significant improvement in model fit (*Gestation*^{2}, Wald $\chi^2 = 1.53$, $p = 0.22$).

The use of fractional polynomials may provide some insight into the function by giving a wider range of possible shapes (Royston *et al,* 1999). Such an approach attempts to find the best power transformation, usually from the candidate powers –2, -1, -0.5, 0, 0.5, 1, 2, 3, where x^0 denotes $log_e x$. More than one term (degree) representing any variable can be included in the model. The powers are compared and the final model selected using changes in deviance. Fractional polynomials are straightforward to apply using the `FRACPOLY` function in STATA. However, there was no statistical evidence of an improvement in model fit, compared to the linear model, from non-linear models of degree 1 ($p = 0.171$) nor from moving to models of degree 2 ($p = 0.738$).

Figure G.2 Observed and Modelled Logit by Gestational Age at Birth



The linear relationship between gestational age and in-unit mortality was given by:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_G . gest_i$$

$\hat{\beta}_0 = 16.16$ (s.e. 0.88)

$\hat{\beta}_G = -0.66$ (s.e. 0.03)

where: *gest* = gestational age at birth in completed weeks.

There was, therefore, strong evidence for a negative relationship between gestational age at birth and in-unit mortality. Expressed as an odds ratio the estimated reduction in odds of in-unit death for each additional week of gestational age was 0.52 (95% CI: 0.48 to 0.55); $p < 0.0001$.

The discriminatory ability of the model as measured by the area under the ROC curve (A_{ROC}) was 0.881 (§6.3.1), and there was no evidence of poor calibration from the Hosmer & Lemeshow goodness-of-fit test: $\hat{C} = 3.70 \sim \chi^2_5$, p = 0.59 (§6.3.2).

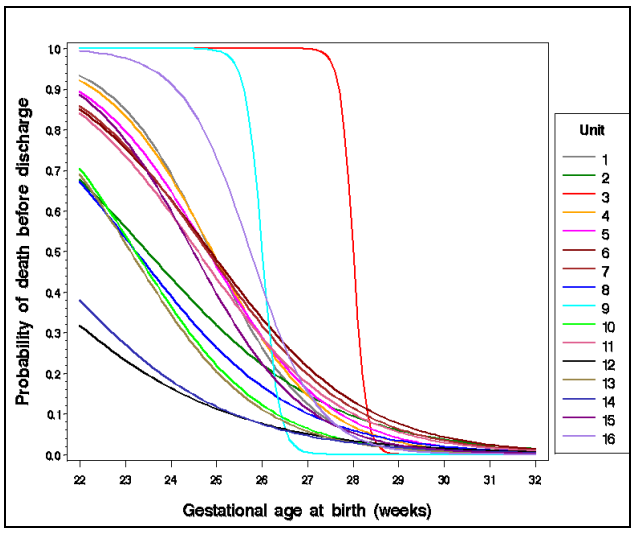
The introduction of an interaction term between gestational age and NICU into the model was used to investigate the homogeneity of the relationship between NICUs across the region. The estimated odds ratios for mortality for the NICUs are given in Table G.2. Although these varied, some were based on very small sample sizes and have large estimated standard errors.

Table G.2 Log Odds Ratios for mortality for one week increase in gestational age at birth, by NICU

Unit	$\hat{g}$	(s.e.)
1	-0.92	(0.17)
2	-0.50	(0.08)
3	-8.00	(31.49)
4	-0.84	(0.25)
5	-0.76	(0.11)
6	-0.60	(0.08)
7	-0.64	(0.10)
8	-0.58	(0.17)
9	-6.39	(24.51)
10	-0.71	(0.24)
11	-0.64	(0.08)
12	-0.43	(0.19)
13	-0.72	(0.27)
14	-0.51	(0.32)
15	-0.82	(0.22)
16	-0.55	(0.20)

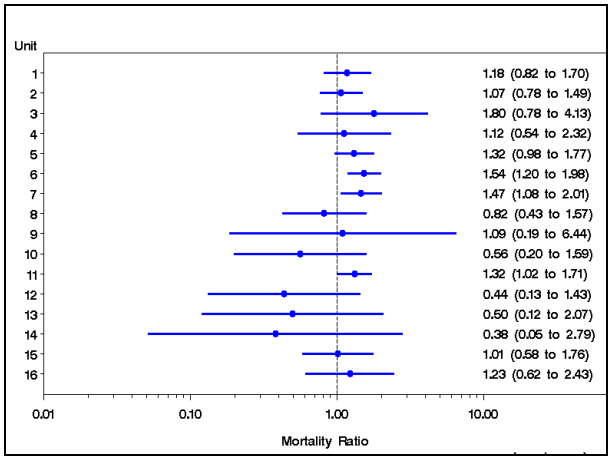
The estimated functions are shown in Figure G.3. Despite these observed differences, there was no statistical evidence that the relationship between gestational age at birth and death before discharge for NICU differed by unit: $\chi^2_{df=15}$=10.19; p = 0.81.

Figure G.3 Estimated probability of death by gestational age at birth by unit



The estimated standardized mortality ratios and 95% confidence intervals, after adjusting for gestational age, are shown in Figure G.4. Three units had estimated 95% confidence intervals wholly above the value one: Units 6, 7 and 11. There were no units with confidence intervals completely below the value one.

Figure G.4 Estimated standardized mortality ratios adjusted for gestational age at birth



Appendix G.2 Sex

There has been shown to be a difference in short-term mortality between newborn boys and girls, with boys showing a higher risk of death. In the United Kingdom during 2001 the 28-day death rate for boys was 4.0 per 1,000 live births and 3.3 for girls (Office of National Statistics, 2003b). This difference has also been show to exist amongst preterm infants (Effer *et al,* 2002; Larroque *et al,* 2004) and cohorts defined by low birth weight (Stevenson *et al,* 2000; Shankaran *et al,* 2002; Italian Collaborative Group on Preterm Delivery, 1988). It has also been shown that male preterm infants receive higher levels of early medical intervention (Elsmén *et al,* 2004).

Data from Sweden show an excess of male infants among preterm deliveries (Ingemarsson, 2003). This was also seen in the TNS data, where 1667 (55%) of the admitted infants were male. There was one observation of unknown sex and this infant was removed from this analysis.

Table G.3 Mortality by sex

	Died		Survived		Total
	n	%	n	%	
Male	1509	90.5	158	9.5	1667
Female	1230	90.6	127	9.4	1357
Total	2739	90.6	285	9.4	3024

In the TNS data the in-unit mortality rates were very similar between the sexes:

odds ratio (male vs. female) = 1.01 (95% CI: 0.79 to 1.30); p = 0.91.

After adding gestational age at birth to the model, there was an increase in the estimated odds ratio. However, there was still no evidence that the value of the odds ratio differed from unity:

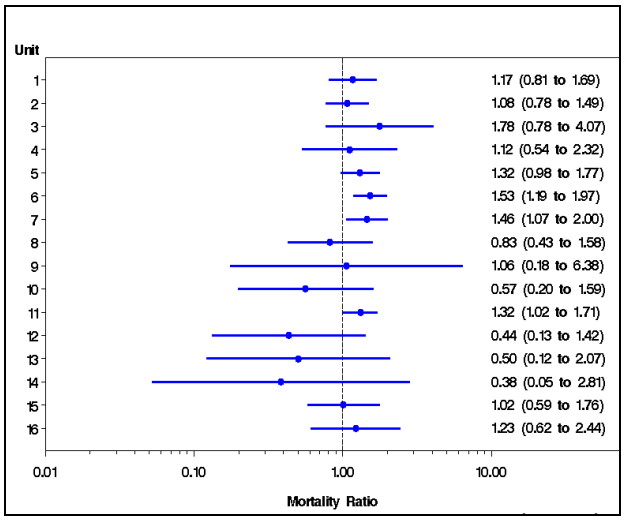
odds ratio (male vs. female) = 1.18 (95% CI: 0.88 to 1.58); p = 0.27

(A_{ROC} = 0.882: $\hat{C}$ = 5.43 ~ χ^2_7, p = 0.61)

There was also no evidence for a gestational age-by-sex interaction: p = 0.21, suggesting that there was a constant difference, on the logit scale, in mortality risk between the sexes. The addition of an interaction term into the logistic model showed no evidence that the odds ratios varied across the neonatal units; p = 0.93. Figure G.5 shows the estimated standardized

mortality ratios after adjusting for sex and gestational age at birth. These are very similar to the results from adjusting for gestational age alone (Figure G.4).

Figure G.5 Estimated standardized mortality ratios adjusted for sex and gestational age at birth



Appendix G.3 Birth weight

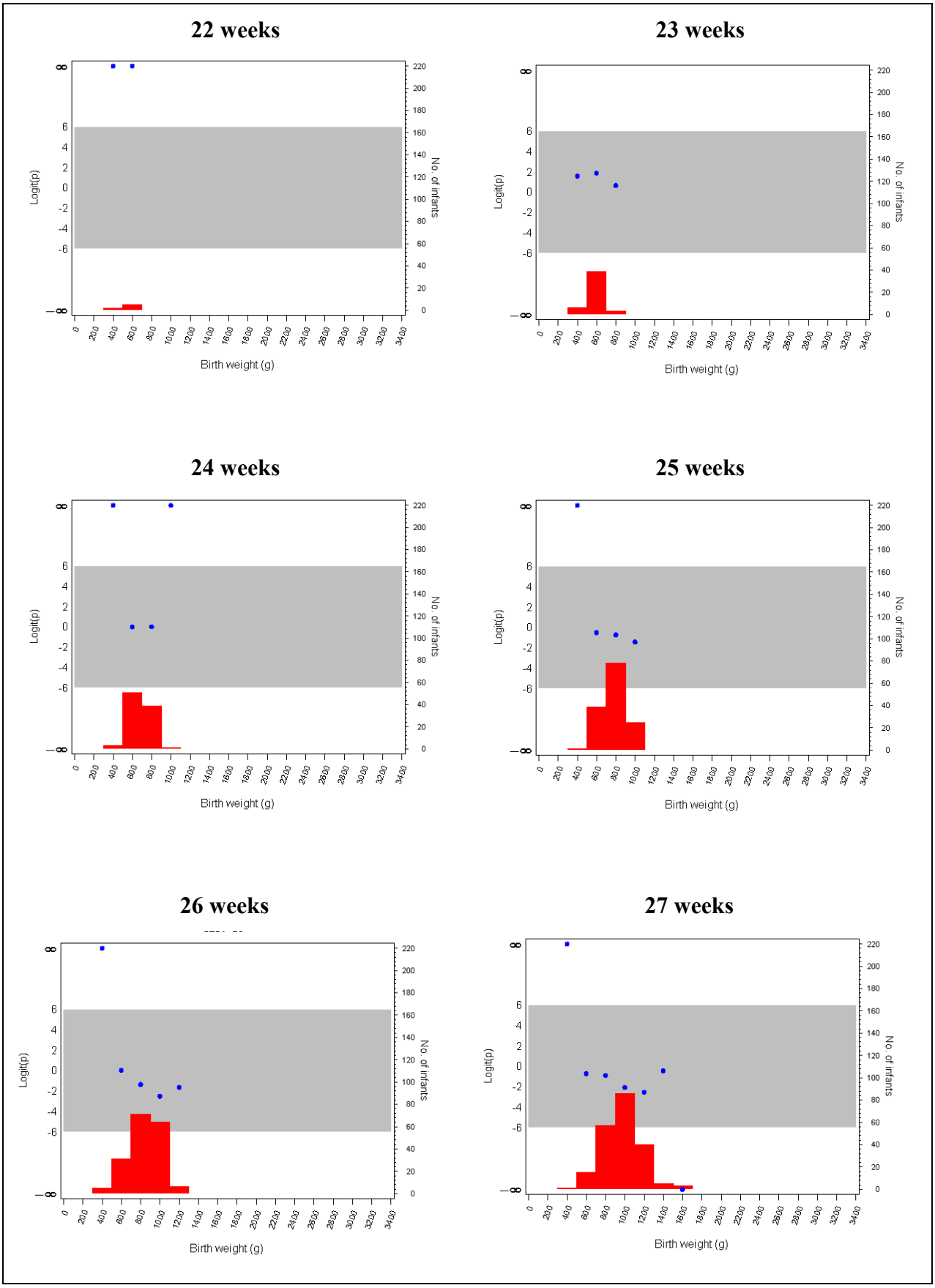
The weight of an infant at birth is known to be associated with its probability of survival (Alberman, 1991). However, it has long been recognised that birth weight in itself is inadequate for predicting mortality (Van Den Berg and Yerushalmy, 1966). Rather, it is the rate of growth in conjunction with gestational age, i.e. birth weight for gestational age, that is more informative (Coory, 1997).

Mortality by birth weight for gestational age

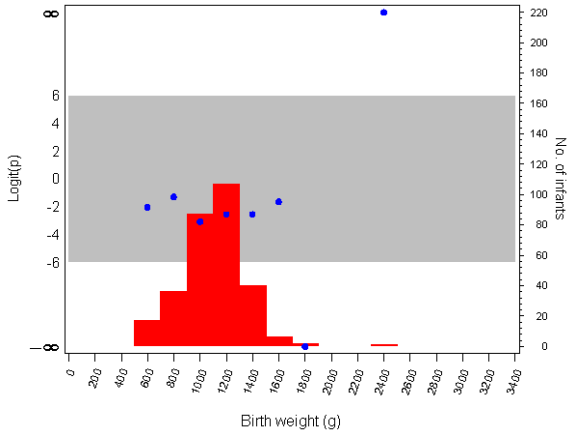
While, for births of all gestational ages, the marginal relationship between birth weight and neonatal mortality is reverse J-shaped (Wilcox and Russell, 1986), it is unclear whether such a relationship still exists conditional on gestational age at birth, particularly for preterm births. While there is plenty of evidence that small for gestational age infants experience poor outcomes, e.g. Regev *et al* (2003), Larroque *et al* (2004), there is conflicting evidence whether being large for gestational age is an indicator for poor prognosis. Some studies have shown evidence of increased mortality with both increasing and decreasing birth weight for gestational age (Draper *et al,* 1999; Yerushalmy, 1970), while other studies have not (Wen *et al,* 2000; McIntire *et al,* 1999).

Inspection of the observed log odds for mortality for the data in this thesis shows little evidence for an increased risk of mortality in large for gestational age infants (Figure G.6).

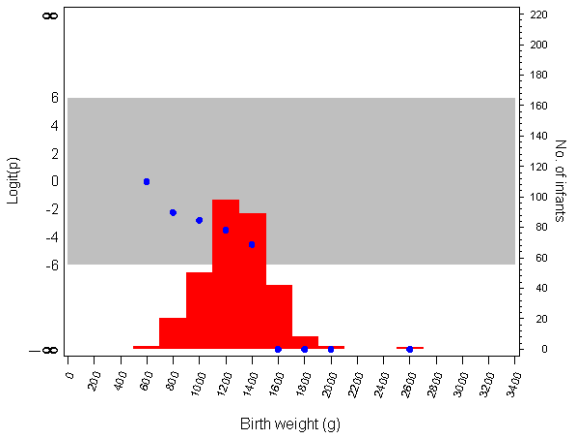
Figure G.6 Observed Logit(death) by weight and gestational age at birth



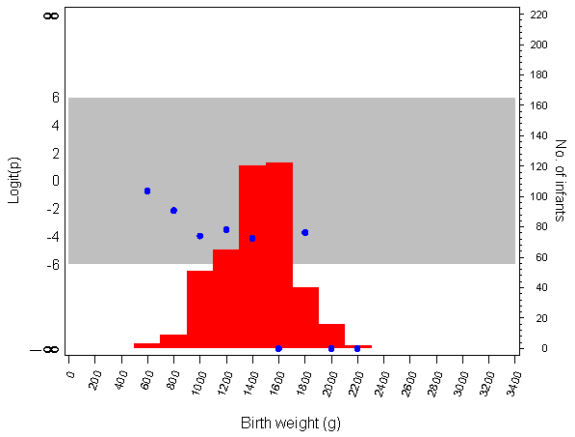
28 weeks



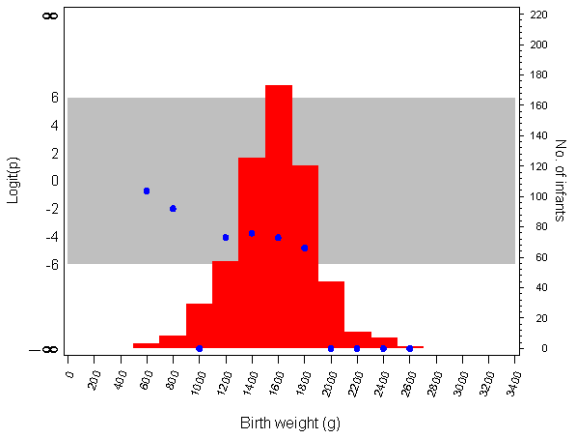
29 weeks



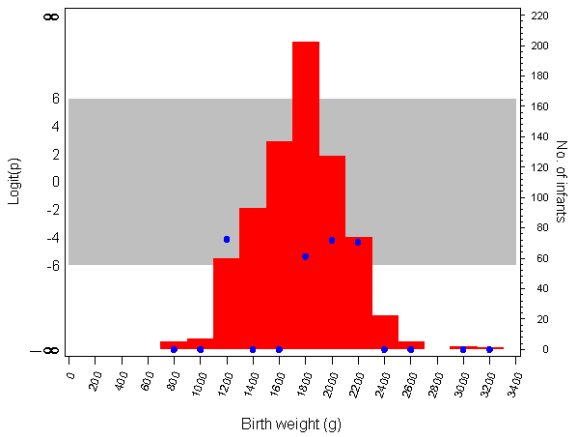
30 weeks



31 weeks



32 weeks



Distribution of birth weights conditional on gestational age

The distribution of birth weight conditional on gestational age at birth is often assumed to follow a Normal distribution (Skjaerven *et al,* 2000; Altman and Chitty, 1997; Kramer *et al,* 2001), although other distributions have been proposed, for example a log-Normal distribution (Oja *et al,* 1991).

Figure G.7 shows the observed birth weights for infants born at 26 and 32 weeks gestational age, together with the estimated Normal and log-Normal probability curves for these data. The two estimated probability distributions are extremely similar. Q-Q plots (Figure G.8) offer evidence that at 32 weeks gestational age the birth weights appear to follow a Normal distribution quite well, apart from a few high values. However, at 26 weeks this is less clear, and there is a suggestion that the data are negatively skewed. This may just be due to random variation, because of a smaller number of observations (177 at 26 weeks and 735 at 32 weeks), but it may also be due to the characteristics of this population. Infants born prematurely are an unrepresentative sample of all fetuses of that gestational age. The very fact that they are born premature means that these infants are likely to be 'unusual' in some way: local data show that less than 2% of all infants born in Leicestershire from 2000 to 2002 were born at 32 weeks or less. There is some evidence that intrauterine growth retardation is an indicator for preterm birth (Ott, 1993; Lackman *et al,* 2001).

Figure G.7 Observed birth weight at 26 and 32 weeks gestational age

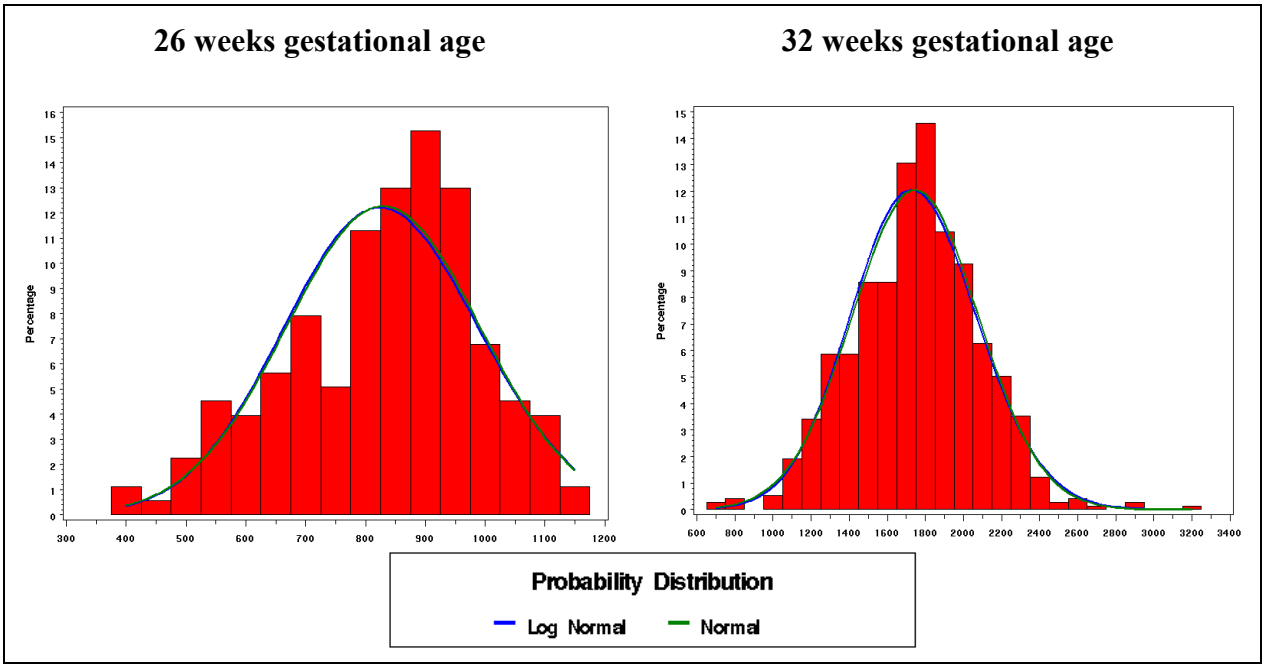
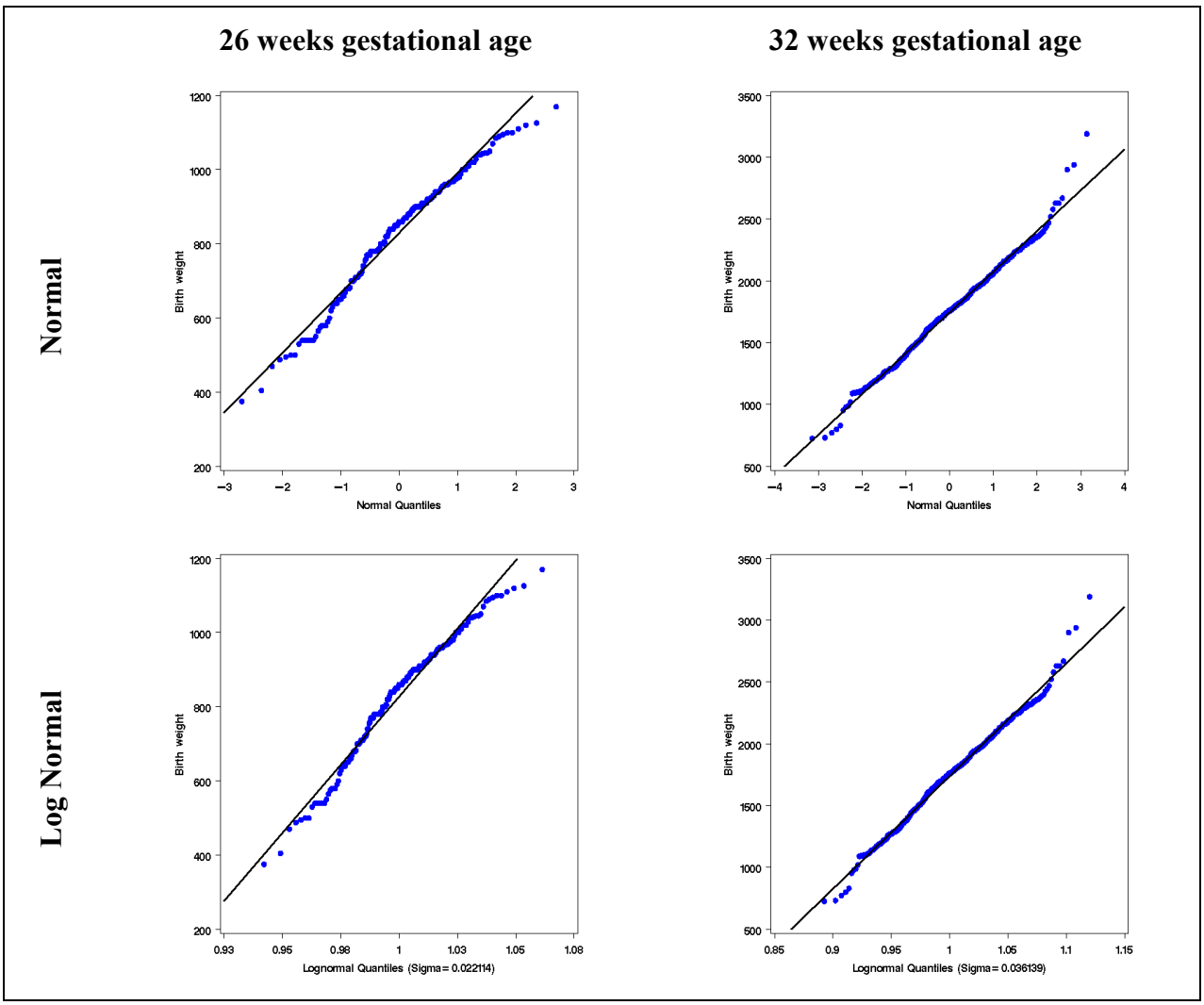


Figure G.8 Normal and LogNormal distribution Q-Q plots for observed birth weights at 26 and 32 weeks gestational age



A further problem arises as it is known that average gestational age specific birth weight differs in various subgroups of infants, for example between sexes (Freeman *et al,* 1995), ethnic groups (Margetts *et al,* 2002), singleton and multiple births (Cohen *et al,* 1997). It may be the case that the relationship between gestational age specific birth weight and mortality also differs between these groups.

Birth weight specific mortality by sex

The data available in this thesis were not sufficient to allow a detailed examination of all possible subgroups. However, there were sufficient data to look at the differences between the sexes. It can be seen from Figure G.9 that the observed mean birth weight by gestational age was consistency lower in girls than boys.

Figure G.9 Observed mean birth weight by gestational age and sex

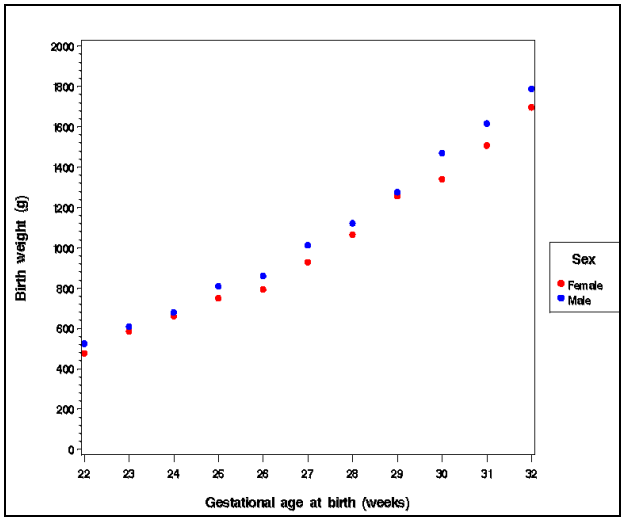


Figure G.10 shows both differences between the sexes in observed birth weight distributions and the observed mortality for those infants born at 26 weeks gestational age. The number of infants is shown in 50g bands, but for the observed mortality 100g bands were used to try to reduce the noise in the data.

There are two points to be noted from Figure G.10. First, both distributions still show signs of left skewness. Second, and more importantly, the difference in observed mortality seems to follow the difference in observed birth weights. This can be seen more clearly in Figure G.11 where the observed birth weight for each observation has been transformed to the difference from the observed mean sex-specific birth weight: 793g for girls and 860g for boys. It can now be seen that the mortality curves are more similar, as are the observed birth weight distributions. This suggests that the differences between the sexes can be accounted for by allowing for the difference in average birth weight, and the analysis needs to take this into account (Wilcox and Russell, 1983; Wilcox and Russell, 1990).

Figure G.10 Observed mortality and number of infants at 26 weeks gestational age by birth weight

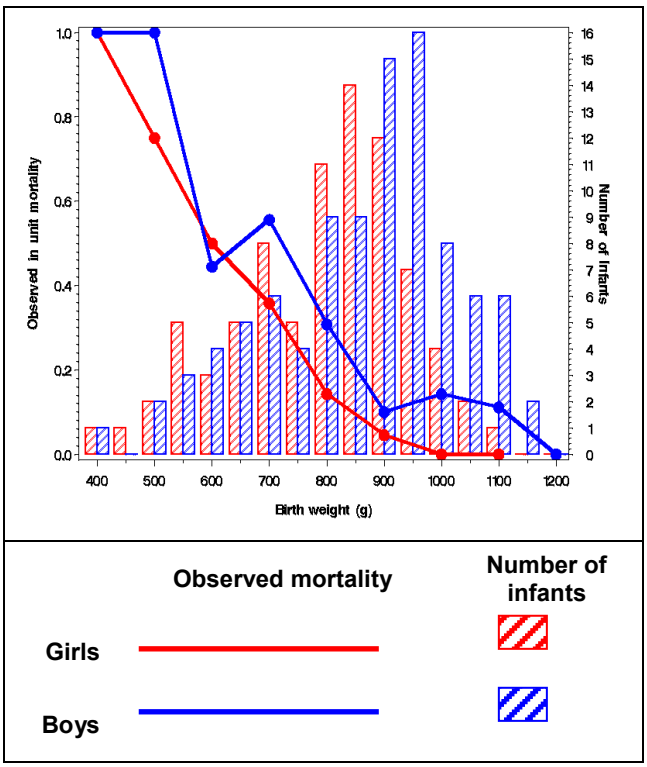
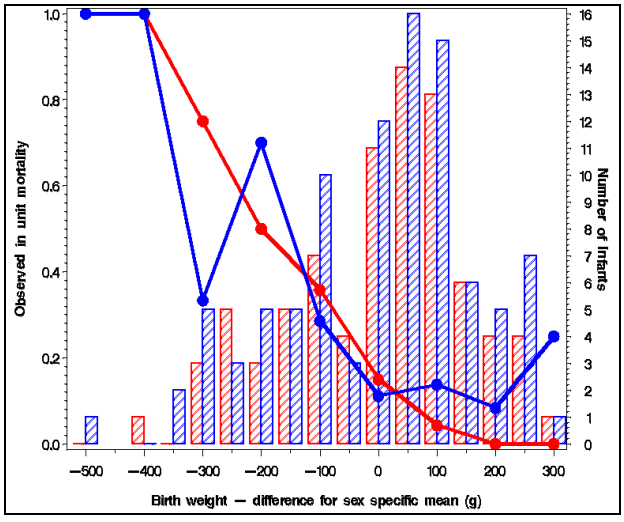


Figure G.11 Observed mortality and number of infants at 26 weeks gestational age by difference from sex-specific mean birth weight



Modelling the relationship between birth weight and mortality

The relationship between weight for gestational age at birth and neonatal mortality can be, and has been, modelled using a number of different approaches. The simplest approach is to include main effects and interactions for birth weight, sex and gestational age into the model.

Alternative approaches, building on the relationship suggested in Figure G.11, involve using the difference between the observed weight and an estimated population mean (or median) birth weight for each gestational age, with such a difference expressed as an absolute difference (Draper *et al,* 1999), z-score or ratio (Kramer *et al,* 1999). Each of these approaches is illustrated below.

There are two other potential methods that were not investigated further here. One was to use percentiles of birth weight. The other method that has been proposed (Van Den Berg and Yerushalmy, 1966) is to first categorise the observations into groups according to their weight at birth, for example in 100g groups (Paneth *et al,* 1982). Then, within each of these groups the observations are ranked according to their gestational age at birth and divided into quartiles. This method thus produces a measure of an infant's gestational age for birth weight. Although this may be a valid way to proceed, in terms of the statistical analysis (Paneth, 1992), it is felt that such an approach confuses the biological relationship between weight and gestational age at birth: *"growth is size for age, not age for size"* (Arnold, 1992) and will not be considered further here.

One further point of note with this model is that both gestational age and birth weight are entered into the model as continuous variables. The models as they exist here do not allow for 'threshold' effects for either variable; for example, a sudden change in the relationship between birth weight for gestational age and mortality at 26 weeks. While such effects cannot be discounted completely, no evidence has been produced for their existence.

The aim was to create a model that was both parsimonious and clinically meaningful, although both of these characteristics may be difficult to achieve together. To investigate the possible approaches, each model (raw birth weight and the three methods looking at deviation from mean birth weight for gestational age: raw difference, z-score and ratio) was specified as a logistic regression model. Using stepwise selection, with 0.10 as the entry and exit threshold level of statistical significance significance, a model was estimated for each approach allowing up to a quadratic term for gestational age, a cubic term for birth weight and all possible interactions. The estimated mean birth weight for gestational age by sex was obtained from a weighted linear regression model (the SAS code for the z-score approach is shown in Appendix D.4 as an example).

The final model for each approach is shown below, together with the estimated area under the ROC-curve ($\boldsymbol{A_{ROC}}$) and the Hosmer & Lemeshow test statistic for lack of fit ($\boldsymbol{C}$). These statistics are measures of the discriminatory ability and the calibration of the model (§6.3).

The models selected and the estimated parameter values are given below.

***Observed birth weight**:*

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_S.sex_i + \hat{\beta}_G.gest_i + \hat{\beta}_W.weight_i + \hat{\beta}_{WW}.weight_i^2$$

$\hat{\beta}_0$ = 12.24 (s.e. 1.13)

$\hat{\beta}_S$ = 0.36 (s.e. 0.16)

$\hat{\beta}_G$ = -0.34 (s.e. 0.05)

$\hat{\beta}_W$ = -0.0081 (s.e. 0.0009)

$\hat{\beta}_{WW}$ = 2.3 x 10^{-6} (s.e. 0.3 x 10^{-6})

(A_{ROC} = 0.897: $\hat{C}$ = 4.35 ~ χ^2_8, p = 0.82)

***Difference from estimated mean birth weight for gestational age**:*

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_G.gest_i + \hat{\beta}_{GG}.gest_i^2 + \hat{\beta}_W.diff_i + \hat{\beta}_{WW}.diff_i^2 + \hat{\beta}_{WG}.diff_i.gest_i + \hat{\beta}_{WWG}.diff_i^2.gest_i$$

$\hat{\beta}_0$ = 47.14 (s.e. 1.0.46)

$\hat{\beta}_G$ = -2.94 (s.e. 0.77)

$\hat{\beta}_{GG}$ = 0.041 (s.e. 0.014)

$\hat{\beta}_W$ = -0.018 (s.e. 0.005)

$\hat{\beta}_{WW}$ = 3.8 x 10^{-5} (s.e. 1.4 x 10^{-5})

$\hat{\beta}_{WG}$ = 5.3 x 10^{-4} (s.e. 0.1 x 10^{-4})

$\hat{\beta}_{WWG}$ = 0.041 (s.e. 0.014)

(A_{ROC} = 0.898: $\hat{C}$ = 5.69 ~ χ^2_8, p = 0.68)

***z-score**:*

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_G.gest_i + \hat{\beta}_{GG}.gest_i^2 + \hat{\beta}_W.zscore_i + \hat{\beta}_{WW}.zscore_i^2$$

$\hat{\beta}_0$ = 33.99 (s.e. 9.49)

$\hat{\beta}_G$ = -1.95 (s.e. 0.70)

$\hat{\beta}_{GG}$ = 0.023 (s.e. 0.013)

$\hat{\beta}_W$ = -0.52 (s.e. 0.07)

$\hat{\beta}_{WW}$ = 0.20 (s.e. 0.03)

(A_{ROC} = 0.897: $\hat{C}$ = 5.83 ~ χ^2_8, p = 0.67)

Ratio:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_G.gest_i + \hat{\beta}_{G^2} gest_i^{\,2} + \hat{\beta}_W.ratio_i + \hat{\beta}_{WW}.ratio_i^{\,2} + \hat{\beta}_{WG}.ratio_i.gest_i$$

$\hat{\beta}_0$ = 57.60 (s.e. 11.77)

$\hat{\beta}_G$ = -2.78 (s.e. 0.77)

$\hat{\beta}_{GG}$ = 0.032 (s.e. 0.013)

$\hat{\beta}_W$ = -21.98 (s.e. 4.99)

$\hat{\beta}_{WW}$ = 5.40 (s.e. 0.91)

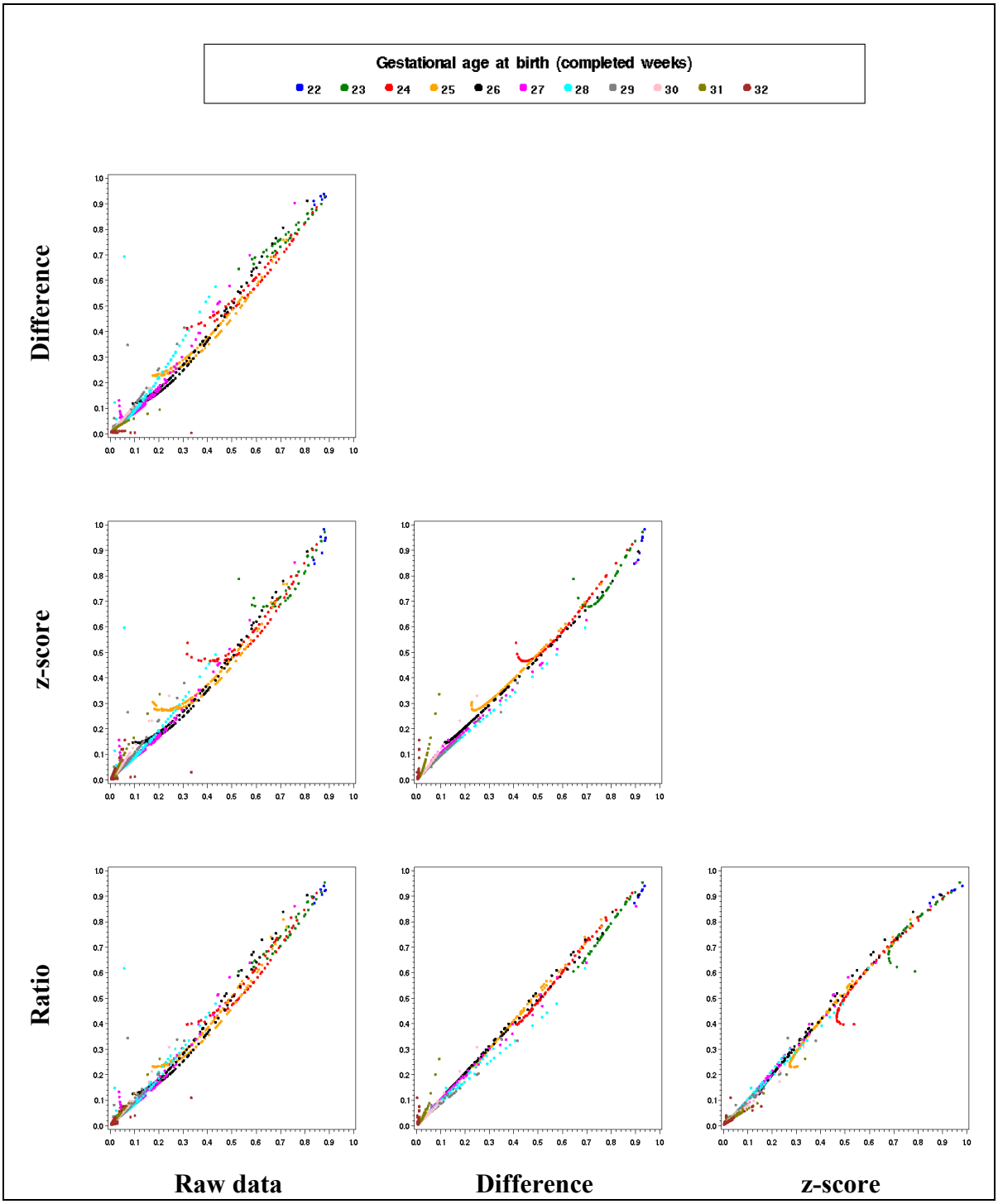
$\hat{\beta}_{WG}$ = 0.30 (s.e. 0.17)

(A_{ROC} = 0.898: $\hat{C}$ = 2.86 ~ χ^2_8, p = 0.94)

Although each model had good discriminatory ability, and there was no evidence of poor calibration, they differed in the number of parameters included in the model and in the predicted probabilities of death before discharge (Figure G.12).

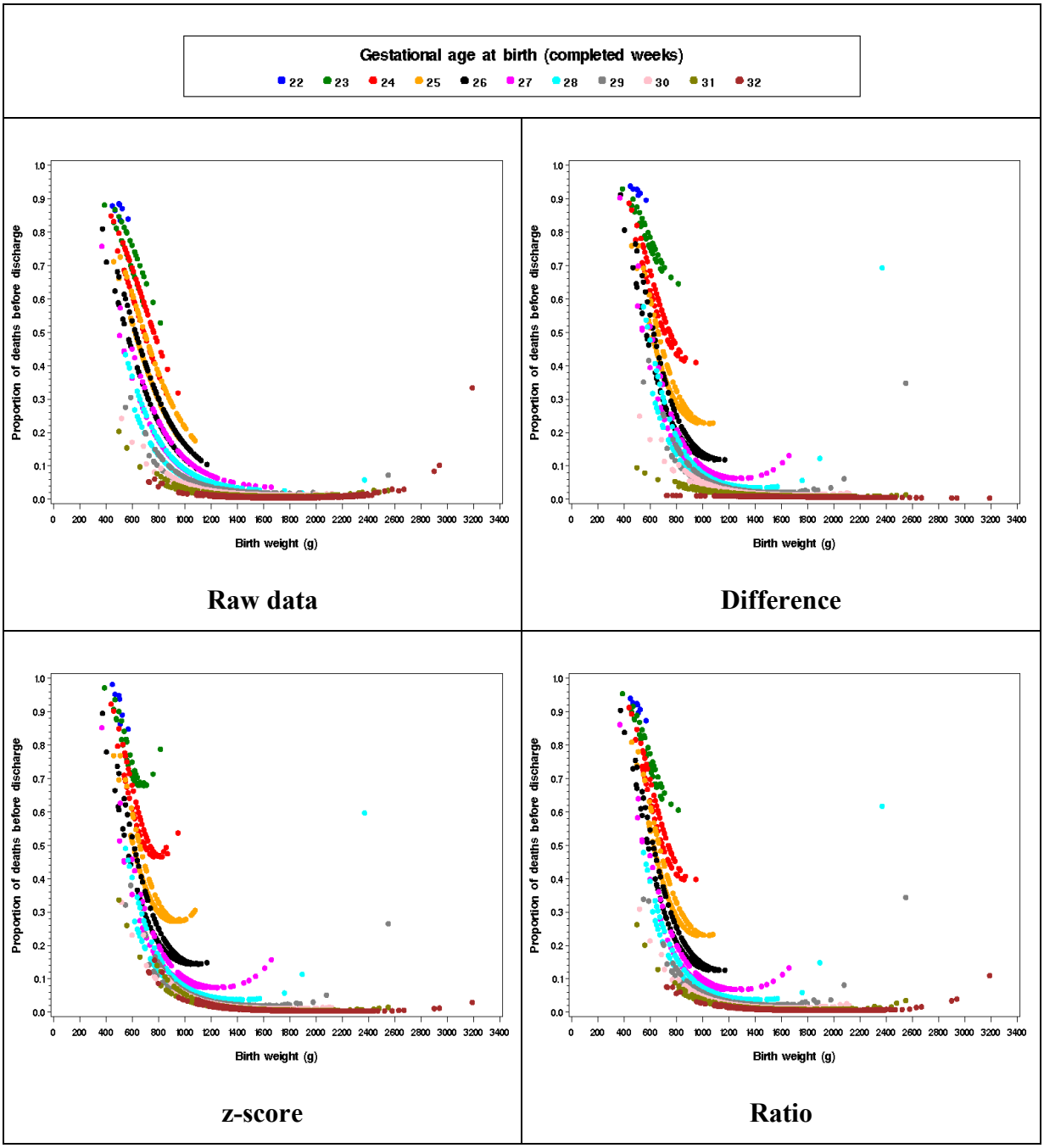
It is unclear which of the models described above is the most appropriate. The model using z-scores showed a reversed J-shaped relationship for all gestational ages, although this is much less marked for 31 and 32 week. Such a strong relationship was not apparent from the observed data. The three other approaches all showed a monotonic, descending, relationship for early gestational ages (≤ 26 weeks) over the range of observed values. There was a reversed J-shaped relationship at higher gestations, although this was not seen for 32 weeks in the *difference* model (Figure G.13). However, despite these differences, the estimated area under the ROC-curve was extremely similar for all of the models, and little different from that estimated from the model with gestational age alone: 0.881.

Figure G.12 Estimated probabilities of death by birth weight for gestational age model



The models using the raw data and the z-score had the least number of terms in the model, with the former having the advantage that it was not necessary to estimate the mean sex-specific birth weight for each gestational age.

Figure G.13 Estimated probability of death by sex, gestational age and birth weight



One observation born at 28 weeks and a weight of 2370g died before discharge. It was of concern that this observation may be unduly influencing the shape of the birth weight function. Although the record values of the deleted observation were check with the original TNS questionnaire and found to be correct, this observation was removed and the four models derived again.

The three models using differences from the estimated gestational age-sex mean (*difference*, *z-score* and *ratio*) remained in the same form, and the values of the parameters estimates

showed only small changes (details not shown). However, the new model using the observed data directly (*raw data*) contained a cubic term for birth weight statistically significant at the 10% significance level (p = 0.060):

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_S.sex_i + \hat{\beta}_G.gest_i + \hat{\beta}_W.weight_i + \hat{\beta}_{WW}.weight_i^2 + \hat{\beta}_{WWW}.weight_i^3$$

$\hat{\beta}_0$ = 15.70 (s.e. 2.10)

$\hat{\beta}_S$ = 0.35 (s.e. 0.16)

$\hat{\beta}_G$ = -0.35 (s.e. 0.05)

$\hat{\beta}_W$ = -0.017 (s.e. 0.005)

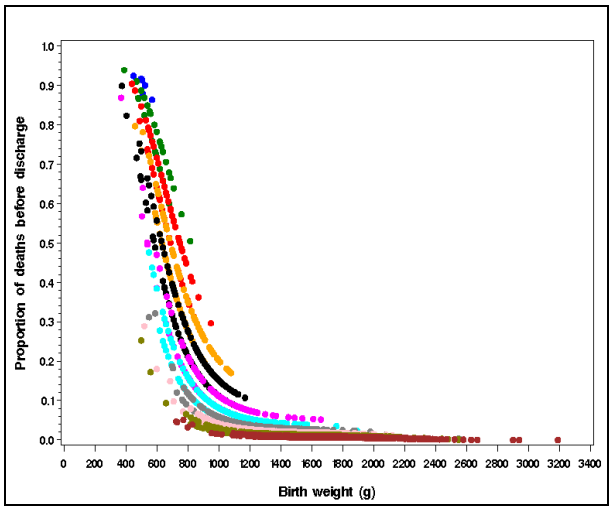
$\hat{\beta}_{WW}$ = 1.0 x 10^{-5} (s.e. 0.4 x 10^{-5})

$\hat{\beta}_{WWW}$ = -2.1 x 10^{-9} (s.e. 1.1 x 10^{-9})

(A_{ROC} = 0.897: $\hat{C}$ = 3.41 ~ χ^2_8, p = 0.91)

The new functions showed no evidence of increased mortality at high birth weight (Figure G.14).

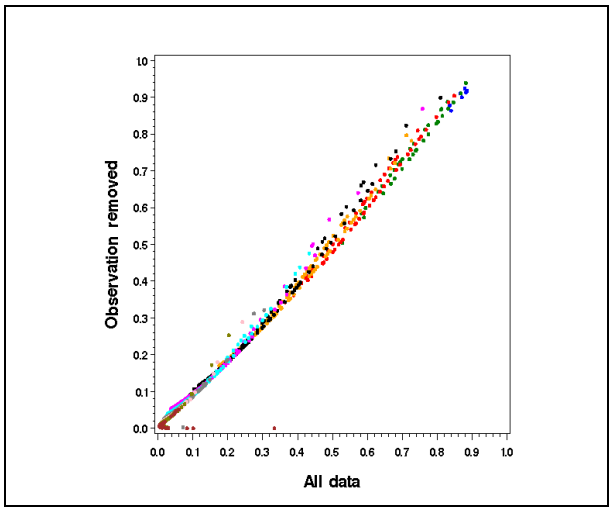
Figure G.14 Probability of death with outlier removed



However, most differences between the two models were small, with only a few observations with high gestational age and birth weight changing their predicted probabilities of death to any great extent (Figure G.15). As this model was not being used to obtain individual predicted probabilities, and the observation was believed to have been recorded correctly, the observation was included in all further analyses. However, it is acknowledged that there is

still uncertainty in the predicted probability of death for infants born at 32 weeks gestational age at over 2500g.

Figure G.15 Comparison of predicted probabilities, with and without outlier



The use of fractional polynomials may provide some insight into the functions by giving a wider range of possible shapes. The approach taken in this thesis was to investigate models up to second-degree fractional polynomials (that is, up to two terms for each variable) using an iterative process of first determining the optimum polynomials for each main effect and then investigating the interactions. The process was repeated until a stable model is found, using the change in deviance.

Using such an approach none of the interactions were statistically significant at the 10% significance level and the final model was:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_S.sex_i + \hat{\beta}_G.gest_i^{-2} + \hat{\beta}_W.weight_i^{-1}$$

$\hat{\beta}_0$ = -9.79 (s.e. 0.48)

$\hat{\beta}_S$ = 0.36 (s.e. 0.16)

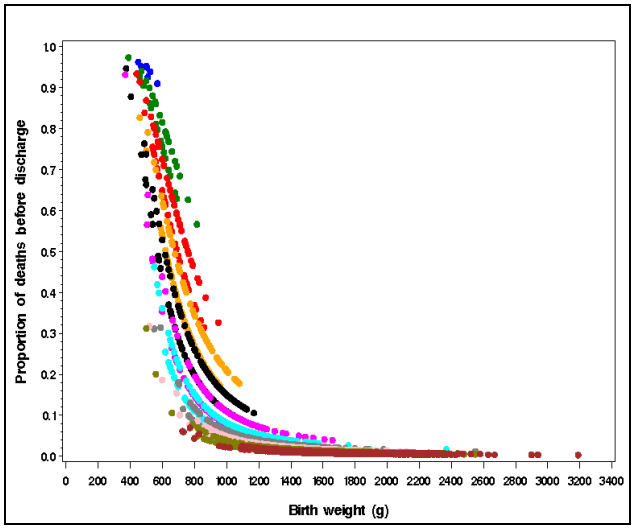
$\hat{\beta}_G$ = 3341.91 (s.e. 439.53)

$\hat{\beta}_W$ = 2763.44 (s.e. 299.56)

(A_{ROC} = 0.897: $\hat{C}$ = 6.10 ~ χ^2_8, p = 0.64)

The estimated probabilities of death from this model are shown in Figure G.16.

Figure G.16 Estimated probability of death (fractional polynomial model)

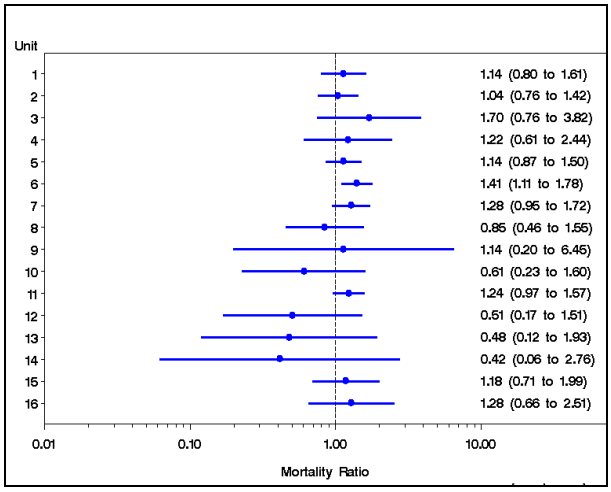


The functions estimated using fractional polynomials (Figure G.16) appear to most closely match the cubic model using the observed values directly (Figure G.14). In reality, there was little difference between all five models considered. The discriminatory abilities, as measured by the area under the ROC-curve, are very close: ranging from 0.897 to 0.898 (compared to 0.881 for gestational age alone). Although detailed model checking was not carried out, no model showed evidence of poor fit using the Hosmer and Lemeshow goodness-of-fit test. The differences between the models in predicted values occurred for the infants with high birth weights for gestational age. The fractional polynomial model was the simplest model considered, having only one term for each variable. In addition, this model and the model using the observed values directly (*raw data*) it did not require the estimation of mean sex-specific birth weights for each gestational age. For the other two approaches, the variable representing the infants' sex was not included in the final model. As a consequence, the final models did not contain the full model uncertainty: the uncertainty in the sex and gestational age-specific birth weights was ignored.

The fractional polynomial model was used in this Section, but when more complex models are investigated later in the Thesis, gestational age and birth weight will be included using the *raw data* approach to allow the easier introduction of interactions with other variables.

Using the fractional polynomial model, there was no evidence that the relationship between birth weight and in-unit mortality differed between the units: $p = 0.60$ for an interaction between the inverse of birth weight and unit of care. The standardized mortality ratios obtained using this model are shown in Figure G.17. Only one unit (Unit 6) had an estimated confidence interval wholly greater than unity.

Figure G.17 Estimated standardized mortality ratios adjusted for sex, birth weight and gestational age at birth



Appendix G.4 APGAR score

The Apgar score was originally derived as a simple neonatal morbidity scoring system (Apgar, 1953) and was described in §4.4.8. There is evidence for an association between low Apgar score and increased mortality, including in preterm infants (Casey *et al,* 2001; Weinberger *et al,* 2000). Apgar scores are usually derived at two time points: one minute after birth and again at five minutes. A low Apgar score represents high morbidity and the rate of mortality declines with increasing Apgar scores

In these TNS data the values of the infants' Apgar scores generally (and unsurprisingly) rose from the first to the second evaluation (Table G.4). There are, however, some scores that fell over this time period, and while these may indicate interesting pathology, the small number of such observations means that they are unlikely to be contributing to risk-adjustment as required here. While change in Apgar score may be of prognostic use, it is likely to be influenced by early neonatal care. This means that it is unsuitable to be included in a model to investigate the quality of care. The same argument also holds for Apgar score at five minutes. For this reason, only the Apgar score at one minute was considered further.

Table G.4 Number of infants by Apgar scores at 1 and 5 minutes after birth

		Apgar Scores at 5 minutes											
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Apgar Scores at 1 minute	**0**	4	1	1	3	2	2	2	2	1	0	1	**19**
	1	1	7	7	4	8	11	14	8	9	8	1	**78**
	2	0	1	11	6	8	18	21	15	24	12	4	**120**
	3	0	1	1	8	14	15	31	37	35	31	2	**175**
	4	0	0	0	1	3	10	23	40	46	58	14	**195**
	5	0	0	1	0	3	11	16	50	65	96	30	**272**
	6	0	1	0	0	0	2	11	36	113	141	31	**335**
	7	0	0	0	0	2	2	4	16	71	208	46	**349**
	8	0	1	0	0	0	3	0	5	52	351	93	**505**
	9	0	1	0	0	1	1	1	2	5	402	328	**741**
	10	0	0	0	0	0	0	0	1	1	1	35	**38**
Total		**5**	**13**	**21**	**22**	**41**	**75**	**123**	**212**	**422**	**1308**	**585**	**2827**

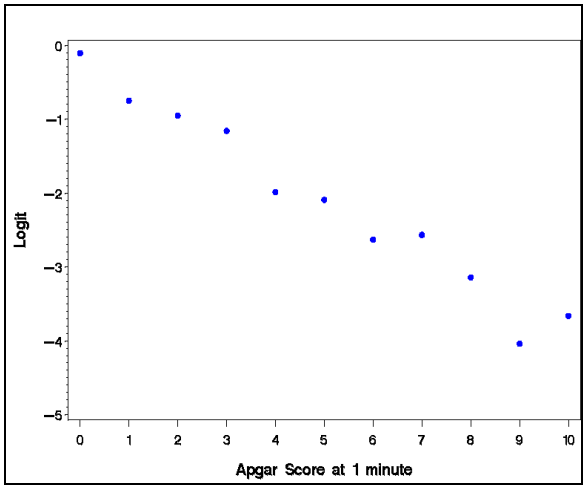
In these data the Apgar score was not obtained at one minute for 139 infants (4.6%). These observations were excluded from the analysis. As expected, for the remaining observations mortality rates fell with increasing Apgar Score (Table G.5).

Table G.5 Mortality by Apgar score at 1 minute

	Apgar Score										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Infants	19	81	122	180	199	282	342	364	507	750	40
Died	9	26	34	43	24	31	23	26	21	13	1
% died	47.4	32.1	27.9	23.9	12.1	11.0	6.7	7.1	4.1	1.7	2.5

When Apgar score was included as a continuous variable in a logistic regression model there was strong evidence of a linear relationship with death before discharge (odds ratio for 1 point increase in Apgar score = 0.68; 95% CI 0.64 to 0.72, $p < 0.0001$), with no evidence for a nonlinear relationship ($p = 0.56$) (Figure G.18).

Figure G.18 Log Odds of death by Apgar Score at 1 minute



There was evidence of an interaction between Apgar score at one minute and gestational age at birth:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_A.apgar1_i + \hat{\beta}_G.gest_i + \hat{\beta}_{GA}.gest_i.apgar1_i$$

$\hat{\beta}_0 = 11.29$ (s.e. 2.06)

$\hat{\beta}_A = 0.70$ (s.e. 0.39)

$\hat{\beta}_G = -0.44$ (s.e. 0.08)

$\hat{\beta}_{GA} = -0.034$ (s.e. 0.015)

where: *apgar1* = Apgar Score at 1 minute

gest = gestational age at birth in completed weeks.

($A_{ROC} = 0.900$: $\hat{C} = 15.99 \sim \chi^2_8$, p = 0.043)

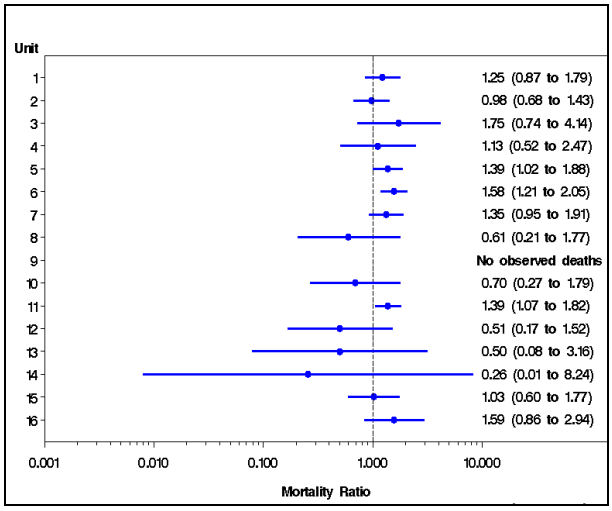
The value of the *C*-statistic suggests that there is some evidence that this model is a poor fit to the data, but there was no evidence for a non-linear relationship. There was also no evidence of different relationships between Apgar score at one minute and outcome between the units (p = 0.90).

When the model with Apgar score at one minute and gestational age was used to indirectly standardize the in-unit mortality, there was a problem with Unit 9. The only observed death at this unit had a missing Apgar score and was excluded from the model. As there were then no observed deaths for Unit 9, there was quasi-complete separation of the data and the estimates became unstable and had large estimated standard errors. To solve this problem

Unit 9 was excluded from the analysis. Alternative approaches could have been adopted, such as imputation of the missing value or by using a different model (e.g. a weighted logistic regression model

When the SMRs were estimated for the other units (Figure G.19), Units 5 and 6 had estimated confidence intervals that did not contain unity.

Figure G.19 Estimated standardized mortality ratios adjusted for Apgar score at one minute and gestational age at birth



Appendix G.5 Ethnic origin

The relationship between ethnicity and neonatal mortality is unclear. There is evidence from the USA that infants born to black mothers experience higher rates of mortality than those born to white mothers (Iyasu *et al,* 2002). However, black infants have higher rates of preterm delivery and further evidence from the USA suggests that after adjustment for gestational age, or birth weight, black preterm infants experience lower mortality that white infants (Cooper *et al,* 1993; Singh *et al,* 1997) and have lower morbidity (Berman *et al*, 2001). There is some evidence of a similar phenomenon in the UK with higher neonatal mortality rates overall in Asian and West Indian populations, but not for preterm infants (Singh *et al,* 1997; Berman *et al,* 2001; Iyasu *et al,* 2002).

In this thesis, the Asian group has been relabelled as 'South Asian' to emphasise the fact that those categorised as Asian are from families originating in South Asia, more particularly from the Indian sub-continent. Other Asian groups, such as Chinese or Filipino, are categorised by TNS as 'Other'.

Table G.6 Mortality by ethnic group of infant

Ethnic group	n	died	%
European	2551	225	8.8
South Asian	239	34	14.2
Other/unknown	235	26	11.1

There was evidence for a difference in mortality between infants of 'European' ethnic origin and those of 'Asian' ethnic origin: overall p = 0.017. However, after the inclusion of gestational age in the logistic regression model, there was no longer evidence for a difference between the ethnic groups: p = 0.19 (Table G.7).

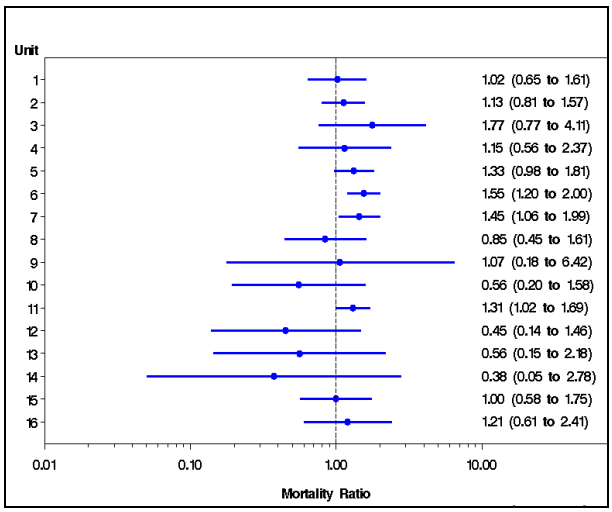
Table G.7 Odds ratios for mortality by ethnic group

Ethnic group	Odds ratio	(95% CI)	p-value
Unadjusted for gestational age			
European	reference		
South Asian	1.72	(1.16 to 2.53)	0.0064
Other/unknown	1.29	(0.83 to 1.98)	0.25
Adjusted for gestational age			
European	reference		
South Asian	1.35	(0.82 to 2.20)	0.23
Other/unknown	0.72	(0.43 to 1.20)	0.21
(A_{ROC} = 0.882: $\hat{C}$ = 5.43 ~ χ^2_7, p = 0.61)			

There was no evidence for an interaction with gestational age (p = 0.85), nor for differences in the relationship between the neonatal units (p = 0.99).

After adjustment for ethnic group and gestational age, Units 6, 7 and 11 had estimated 95% confidence intervals with lower limits greater than unity.

Figure G.20 Estimated standardized mortality ratios adjusted for ethnic origin and gestational age at birth



Appendix G.6 Congenital anomalies

The presence of infants with congenital anomalies can affect in-unit mortality rates in two ways. First, an increased rate of prenatal diagnosis and pregnancy termination for congenital anomalies is likely to decrease in-unit mortality rates (Liu *et al,* 2002). Second, high rates of admissions of infants with congenital anomalies are likely to increase in-unit mortality (Sankaran *et al,* 2002). The TNS data allowed an investigation of the latter process but not the former.

Although infants with lethal congenital anomalies have been excluded from all of these analyses, infants with anomalies not thought to be inevitably lethal have been included. One hundred and sixty-nine infants were recorded as having a congenital anomaly: all of which were chromosomal anomalies.

Table G.8 Mortality by presence of congenital malformation

	Survived		Died		Total
	n	%	n	%	
None	2588	90.6	268	9.4	2856
Malformation	152	89.9	17	10.1	169
Total	2740	90.6	285	9.4	3025

The unadjusted odds ratio of morality was 1.08 (95% CI: 0.64 to 1.82); p = 0.77. With the introduction of gestational age into the model, there was evidence for a quadratic relationship

between gestational age and mortality and for interactions between the presence of a congenital malformation and both linear and quadratic terms for gestational age:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_C.conmal_i + \hat{\beta}_G.gest_i + \hat{\beta}_{CG}.conmal_i.gest_i + \hat{\beta}_{GG}.gest_i^{\,2} + \hat{\beta}_{CGG}.conmal_i.gest^2$$

$\hat{\beta}_0 = 58.62$ (s.e. 41.41)

$\hat{\beta}_C = 74.39$ (s.e. 37.65)

$\hat{\beta}_G = 5.15$ (s.e. 2.98)

$\hat{\beta}_{CG} = -5.74$ (s.e. 2.69)

$\hat{\beta}_{GG} = -0.11$ (s.e. 0.05)

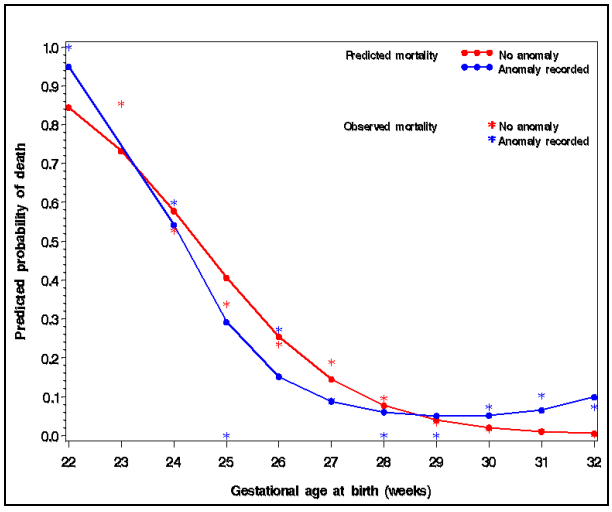
$\hat{\beta}_{CGG} = 0.11$ (s.e. 0.05)

where: $conmal = \begin{cases} 1 & \text{if malformation present} \\ 0 & \text{if no malformation present} \end{cases}$

gest = gestational age at birth in completed weeks.

($A_{ROC} = 0.891$: $\hat{C} = 7.15 \sim \chi^2_5$, p = 0.21)

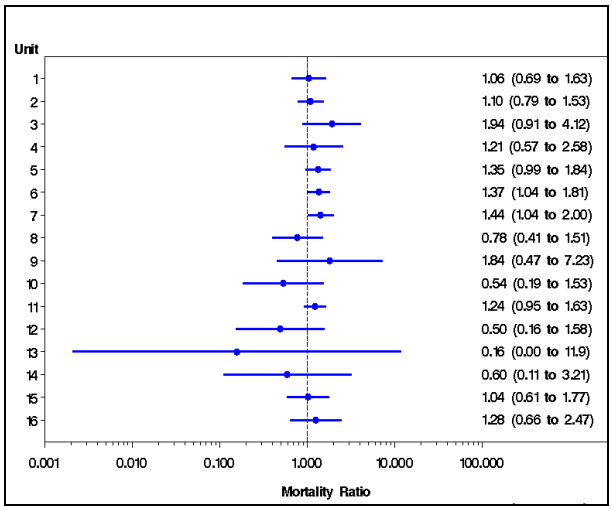
Figure G.21 Observed and estimated probability of death by presence of a congenital malformation and gestational age at birth



It is unclear what interpretation can be put on such a relationship and potential reasons will not be pursued in this thesis. One explanation may be varying incidence rates of different types of congenital malformations with increasing gestational age, but data are not available from TNS to explore this further. The addition of an interaction term into the logistic model

showed no statistical evidence that the odds ratio varied across the neonatal units (p = 0.99). After adjustment, Units 6 and 7 had estimated confidence intervals that did not contain unity.

Figure G.22 Estimated standardized mortality ratios adjusted for congenital malformation and gestational age at birth



Appendix G.7 Base excess

The base excess is the theoretical amount of acid that needs to be given to correct the blood pH: specifically, to titrate one litre of blood to pH 7.4 at a Pco2 of 5.3 kPa (Gray *et al,* 1985:41). A positive value indicates metabolic alkalosis and negative values indicate metabolic acidosis, with the normal range for newborn infants being –10 to –2 mmol/L (Tietz, 1986:1815). For TNS, the maximum base excess in the first 12 hours of life is recorded. Abnormal base excess has been shown to be associated with neonatal mortality in preterm infants (The International Neonatal Network, 1993; Maier *et al,* 1997; Garcia *et al,* 2000; Parry *et al,* 2003b).

Analysis with observations with missing values excluded

Seven hundred and forty eight infants (24.7%) had missing values for base excess, of whom 17 (2.3%) died. For the remaining 2277 infants, the median value for base excess was –5.8mmol/L (mean = -6.5, minimum = -29.5, maximum = 20.1).

Using the observations with known base excess, the unadjusted odds ratio of mortality for each unit increase in base excess was 0.83 (95% CI: 0.80 to 0.85); $p < 0.0001$. There was no evidence for a non-linear relationship between base excess and mortality; the p-value for a

quadratic term for base excess being 0.13. When gestational age was added to the model there was evidence for an interaction between base excess and gestational age ($p < 0.001$).

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_B.baseexcess_i + \hat{\beta}_G.gest_i + \hat{\beta}_{BG}.baseexcess_i.gest_i$$

$\hat{\beta}_0 = 19.36$ (s.e. 2.03)

$\hat{\beta}_B = 0.51$ (s.e. 0.17)

$\hat{\beta}_G = -0.82$ (s.e. 0.08)

$\hat{\beta}_{BG} = -0.025$ (s.e. 0.006)

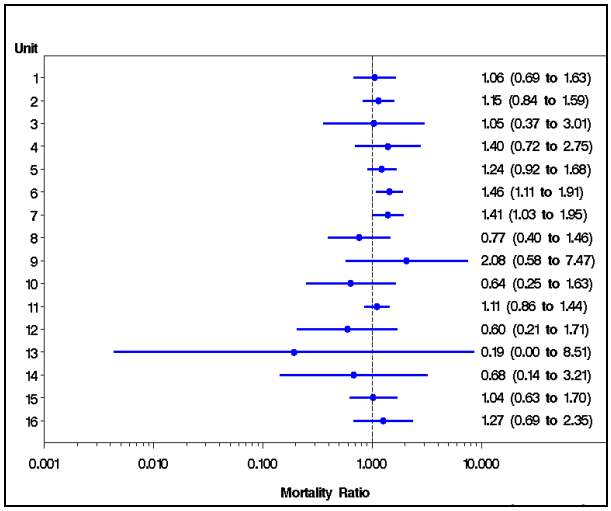
where: *baseexcess* = base excess (mmol/L)

gest = gestational age at birth in completed weeks.

($A_{ROC} = 0.896$: $\hat{C} = 14.08 \sim \chi^2_8$, $p = 0.080$)

There was no evidence of a different relationship amongst the units: $p = 0.96$.

Figure G.23 Estimated standardized mortality ratio adjusted for recorded base excess and gestational age at birth



Analysis with observations with missing values included

The analysis above excluded 748 observations without recorded maximum base excess. The Royal College of Obstetricians and Gynaecologists and the Royal College of Midwives issued a joint recommendation that consideration should be given to the routine measurement of cord blood gasses, which would allow the calculation of base excess (RCOG & RCM, 1999:22). However, a recent survey of obstetric units found that only 54% were following these recommendations (Waugh *et al,* 2001). While the true reason measurements are missing from

TNS is unknown, anecdotal evidence suggests that in most cases base excess was not measured when it was felt likely to be in the normal range (Field, D.J.: Personal communication). In this case, it may be appropriate to substitute the missing values with a 'normal' value. Such an approach is commonly used with published risk-adjustment scores, for example PIM (Shann *et al,* 1997), SNAP (Richardson *et al,* 1993), MMPS (Daley *et al,* 1988). That only 2.3% of those with missing base excess values died before discharge suggests that this may be true (Table G.9). Using this assumption, it was possible to categorise all of the observations according to their estimated base excess by putting those with missing values into the 'normal group'. The groups used here were those from the original CRIB score: >-7.0, -7.0 to -9.9, -10.0 to –14.9 and ≤ 15.0 mmol/L. The new CRIB II score uses more categories but it was felt that these might result in small counts for some groups. In addition, a small number of groups meant that missing values were less likely to be allocated to the wrong group.

Table G.9 Deaths by maximum base excess

Max. base excess (mmol/L)	**Total**	**Died**	**(%)**
Missing values	*748*	*17*	*(2.3)*
Known > -7.0	*1412*	*81*	*(5.7)*
> -7.0 [xii]	2160	98	(4.5)
-7.0 to –9.9	473	63	(13.3)
-10.0 to –14.9	276	69	(25.0)
≤ 15.0	116	55	(47.4)

There was evidence of a difference in mortality rates between the groups ($p < 0.0001$), with increasing mortality with decreasing maximum recorded base excess (Table G.9). The inclusion of gestational age showed evidence for an interaction between gestational age and maximum base excess group: $p = 0.0003$:

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_{B2}.baseexcess2_i + \hat{\beta}_{B3}.baseexcess3_i + \hat{\beta}_{B4}.baseexcess4_i + \hat{\beta}_G.gest_i + \hat{\beta}_{B2G}.baseexcess2_i.gest_i + \hat{\beta}_{B3G}.baseexcess3_i.gest_i + \hat{\beta}_{B4G}.baseexcess4_i.gest_i$$

$\hat{\beta}_0 = 18.19$ (s.e. 1.50)

$\hat{\beta}_{B2} = -2.81$ (s.e. 2.63)

[xii] Missing values and known observations < -7.0 mmol/L

$\hat{\beta}_{B3} = -7.16$ (s.e. 2.37)

$\hat{\beta}_{B4} = -7.06$ (s.e. 2.69)

$\hat{\beta}_{G} = -0.76$ (s.e. 0.06)

$\hat{\beta}_{B2G} = 0.12$ (s.e. 0.10)

$\hat{\beta}_{B3G} = 0.32$ (s.e. 0.09)

$\hat{\beta}_{B4G} = 0.135$ (s.e. 0.10)

where: $baseexcess.2 = \begin{cases} 1 & \text{if maximum base excess} = -7.0 \text{ to } -9.9 \\ 0 & \text{else} \end{cases}$

$baseexcess.3 = \begin{cases} 1 & \text{if maximum base excess} = -10.0 \text{ to } -14.9 \\ 0 & \text{else} \end{cases}$

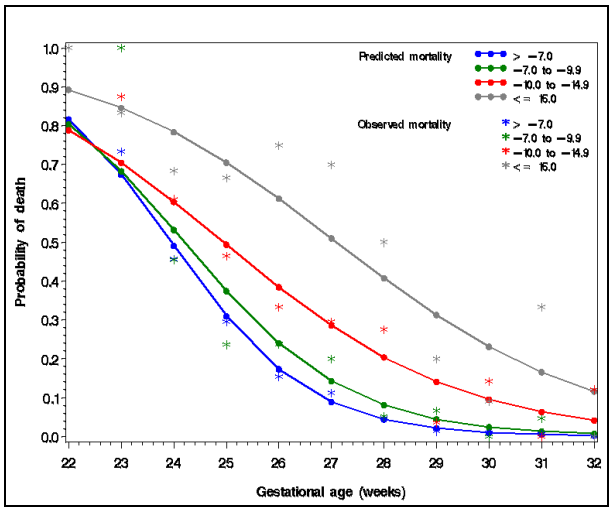
$baseexcess.4 = \begin{cases} 1 & \text{if maximum base excess} \leq 15.0 \\ 0 & \text{else} \end{cases}$

gest = gestational age at birth in completed weeks.

($A_{ROC} = 0.911$: $\hat{C} = 1.10 \sim \chi^2_8$, p = 0.98)

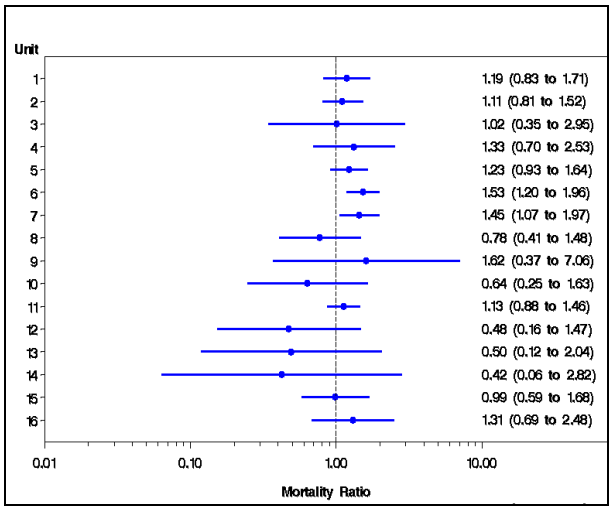
The introduction of interaction terms between the units and base excess groups showed no improvement in the fit of the model: p = 0.98. The estimated functions showed a 'dose response' pattern of higher mortality with decreasing base excess (Figure G.24).

Figure G.24 Observed and estimated probability of death by base excess and gestational age at birth



From the estimated SMRs, there was statistical evidence that Units 6 and 7 had high mortality rates (Figure G.25).

Figure G.25 Estimated standardized mortality ratios adjusted for base excess and gestational age at birth



Appendix G.8 Multiplicity of pregnancy

There has been evidence presented that multiple birth is a risk factor amongst extremely low birth weight infants (501-1000g) (Shankaran *et al,* 2002). However, there is also evidence that twins have better gestational age specific neonatal survival rates than singletons (Kiely, 1998). The TNS records the number of fetuses in each pregnancy, allowing an investigation into the effect of multiple birth on mortality. However, no other details are recorded, so it was not possible to investigate other related potential risk factors in multiple births, such as first-born versus second-born (Shinwell *et al,* 2004; Sheay *et al,* 2004), asynchronous delivery (Livingston *et al,* 2004), monochorionicity and discordant growth (Amaru *et al,* 2004).

The TNS data showed decreasing mortality with increasing multiplicity, although these differences are not statistically significant: overall p-value = 0.23. Since there were a relatively small number of triplets, and only three deaths, a dichotomous variable was used: singleton or multiple birth. Once again there was no statistically significant difference between the groups (Table G.10). However, this approach may be concealing differences by gestational age (Table G.11).

Table G.10 Unadjusted odds ratios for mortality by multiplicity of birth

Multiplicity of birth	Total	Died	(%)	Odds ratio	(95% CI)	p-value
Singleton	2288	225	(9.8)	reference		
Twin	669	57	(8.5)	0.85	(0.63 to 1.16)	0.31
Triplet	68	3	(4.4)	0.42	(0.13 to 1.36)	0.15
Multiple [xiii]	737	60	(8.1)	0.81	(0.60 to 1.10)	0.17

Table G.11 Mortality by multiplicity of pregnancy and gestational age

Multiplicity of birth		Gestational age at birth (weeks)										
		22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
Singleton	**n**	5	38	78	106	138	151	237	241	317	439	538
	died	5	31	40	33	33	29	25	8	7	9	5
	(%)	(100)	(82)	(51)	(31)	(24)	(19)	(11)	(3)	(2)	(2)	(1)
Multiple	**n**	2	10	16	37	39	56	59	71	111	139	197
	died	2	10	10	14	9	9	2	2	1	1	0
	(%)	(100)	(100)	(63)	(38)	(23)	(16)	(3)	(3)	(1)	(1)	(0)
Total	**n**	7	48	94	143	177	207	296	312	428	578	735
	died	7	41	50	47	42	38	27	10	8	10	5
	(%)	(100)	(85)	(53)	(33)	(24)	(18)	(9)	(3)	(2)	(2)	(1)

The effect of gestational age at birth was investigated using a logistic regression model and there was evidence of a gestational age-by-multiplicity interaction (p = 0.0065):

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_M .multiple_i + \hat{\beta}_G .gest_i + \hat{\beta}_{MG} .multiple_i .gest_i$$

$\hat{\beta}_0 = 14.95$ (s.e. 0.95)

$\hat{\beta}_M = 7.19$ (s.e. 2.68)

$\hat{\beta}_G = -0.62$ (s.e. 0.04)

$\hat{\beta}_{MG} = -0.28$ (s.e. 0.10)

where: $multiple = \begin{cases} 1 & \text{if multiple birth} \\ 0 & \text{if singleton birth} \end{cases}$

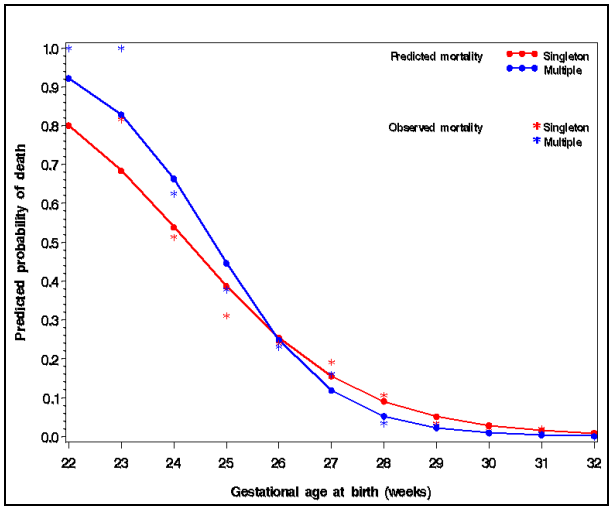
[xiii] Twins and triplets

gest = gestational age at birth in completed weeks

($A_{ROC} = 0.885$: $\hat{C} = 4.97 \sim \chi^2_8$, p = 0.66)

From both the observed mortality (Table G.11), and the model estimates, there appeared to be evidence that infants from multiple pregnancies did worse that singletons if born before about 26 weeks gestational age, but appeared to do better if born after this time.

Figure G.26 Estimated mortality by gestational age and multiplicity of pregnancy



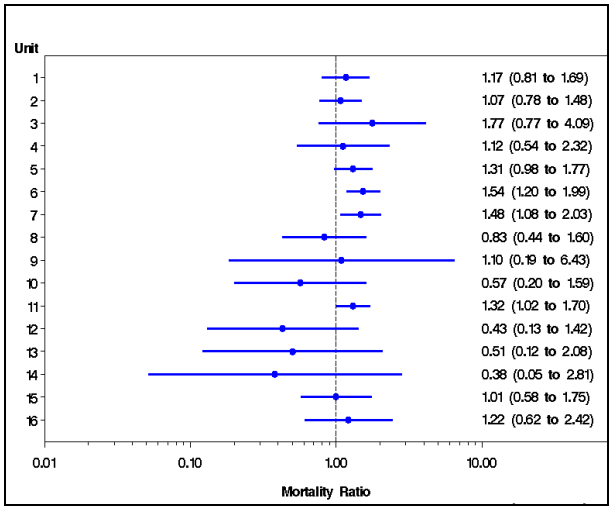
This may help to explain the conflicting evidence discussed earlier. The very small infants, less than 1000g, tend to be born at very early gestational ages (see Figure 2.8): in the TNS data 74% of births below 1000g were at 27 weeks or less. Another study, investigating births from 26 weeks onwards, showed a survival advantage for twins for all gestational ages except at 26 weeks (Kiely, 1998). Such data, together with the finding in this thesis, suggest that there may be a crossing over of neonatal mortality curves at around 26 weeks gestational age.

However, an explanation for this phenomenon is less clear. Intersecting neonatal mortality curves have been observed for infants born at greater gestational ages (Cheung *et al,* 2000). Studying births at 28 weeks gestational age and over, Cheung *et al* observed higher neonatal (and perinatal and infant) mortality for singleton births up to around 37 weeks but lower mortality for singleton for births after this time. Rather than their assumption that *"... twins have better health than singletons initially ..."* such observations may just reflect the fact that twins tend to be delivered earlier than singletons, perhaps through the shortening of the cervix, (Sullivan and Newman, 2004). Singletons are likely to have some pathological cause for their preterm birth other that lack of space in the uterus (Lie, 2000). Whether there was some form of selection bias with the very preterm births observed here is impossible to know:

obstetric care may vary in the early stages of pregnancy, accounting for the differences, but no such variations are known. However, the present work is interested in predicting the outcome of such deliveries to enable appropriate risk-adjustment, rather than an examination of the underlying causes. The reason behind the observed pattern of mortality will not, therefore, be pursued further here.

There was no evidence that the relationship between mortality and multiple birth was different between the units, after adjusting for gestational age ($p = 0.99$). After adjustment for multiplicity and gestational age, Units 6, 7 and 11 had estimated 95% confidence intervals wholly greater than unity.

Figure G.27 Estimated standardized mortality ratios adjusted for multiplicity and gestational age at birth



Appendix G.9 Socio-economic status

The Trent Neonatal Survey does not collect any information directly on the socio-economic status of an infant's family: for example the National Statistics Socio-economic Classifications (NS SEC) or the Standard Occupational Classifications (SOC). However, the postcode of the mother's place of residence was recorded and can be used to investigate any association between area-based socio-economic deprivation and in-unit neonatal morality.

The evidence for an association between socio-economic deprivation and neonatal mortality is equivocal. While some previous studies have shown evidence of an association between neonatal mortality and increased deprivation (Martuzzi *et al,* 1998; Joyce *et al,* 2002; Guildea *et al,* 2001; Joyce *et al,* 2004), it has been suggested that this relationship does not hold for

infants born under 2500g (Leon, 1991; Bambang *et al,* 2000). However, these studies have ecological study designs and, as such, do not show a direct relationship between deprivation and neonatal mortality for an individual and as such may be prone to the ecological fallacy (Selvin, 1958). Other studies that looked at deprivation and neonatal mortality in individuals found no evidence for an association after adjustment for race, sex and gestational age (Paneth *et al,* 1982) nor for gestational age and birth weight (Manktelow and Draper, 2003). The lack of evidence for an association in low birth weight infants, or after adjustment for gestational age, may occur because deprivation is associated with mortality through the increased risk of a preterm, or low birth weight, birth (Aveyard *et al,* 2002; Meis *et al,* 1995; Peacock *et al,* 1995). A Canadian study is currently underway to investigate possible causal pathways to explain the association between socio-economic deprivation and preterm birth at the individual level (Kramer *et al,* 2001).

Several area-based deprivation scoring systems exist that could have been used for these data, for example the Townsend score (Townsend *et al,* 1988), the Jarman score (Jarman, 1983), the Index of Multiple Deprivation (IMD) (Department of the Environment Transport and the Regions, 2000). The measure of socio-economic status chosen was the Index of Multiple Deprivation 2000 (IMD) published by the Department of the Environment, Transport and the Regions. The Department of Social Policy and Social Work at the University of Oxford were commissioned to develop this score to provide an up-to-date measure of deprivation at electoral ward level. Current versions of the Townsend and Jarman scores used data from the 1991 census and, therefore, were more likely to be out-of-date than the IMD which also included more recent data.

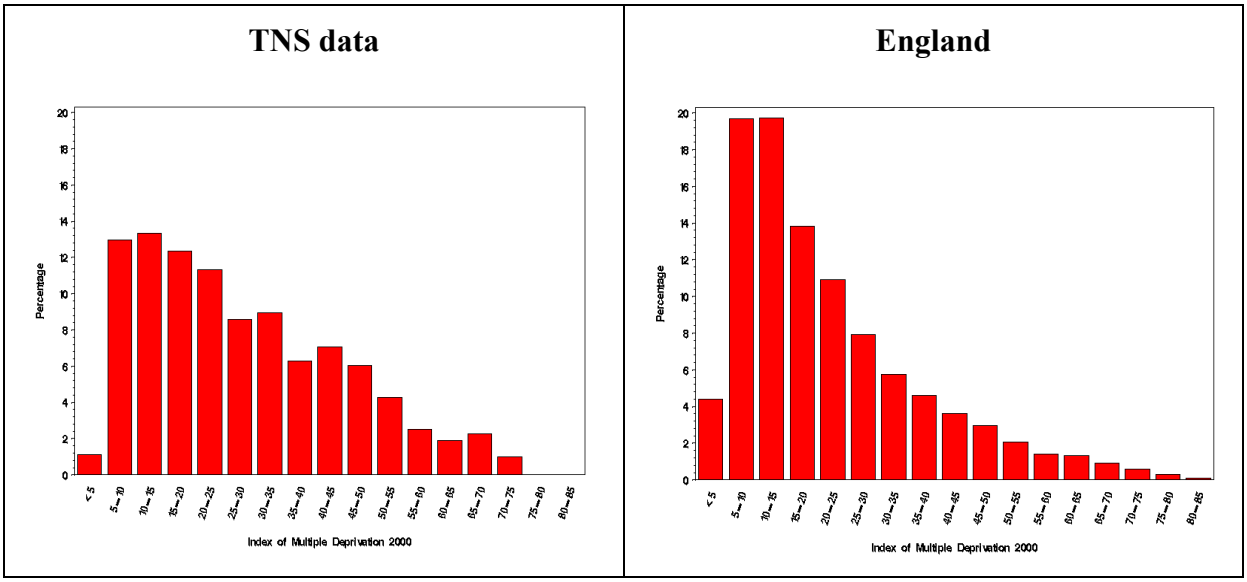
Each electoral ward had an overall IMD 2000 score, which is the sum of six weighted domain scores: income 25%; employment 25%; health & disability 15%; education, skills & training 15%; housing 10%; and geographical access to services 10%. A higher value of the score represents a higher level of socio-economic deprivation.

The IMD 2000 has itself been updated to become the IMD 2004 (Office of the Deputy Prime Minister, 2003). However, there are important differences between IMD 2000 and IMD 2004 that make any direct comparisons, for example to investigate changes in deprivation over time, unfeasible. First, the IMD 2000 'housing' and 'geographical access to services' domains have been replaced with domains called 'barriers to housing and services' and 'living environment', and a new domain of 'crime' has been added. Second, IMD 2000 was electoral ward level statistic whereas IMD 2004 uses smaller Super Output Areas (SOAs).

While the latter difficulty may be partly overcome by averaging IMD 2004 scores over areas, the change in domains means that only the rankings of areas can be investigated rather than their absolute values. Such investigations have found high correlations between the scores (Office of the Deputy Prime Minister, 2003:117-118).

For this thesis, the Postcode Plus® software (AFD Software Ltd, 2004) was used to allocate observations to electoral wards using the full postcode subsectors (e.g. LE1 6TP) of the mother's stated place of residence at the time of the birth. It was not possible to match 106 observations (9 deaths) to their appropriate electoral ward, either because the recorded post code was incorrect or because boundary changes made it impossible to match the post codes to wards with the available software. The 2919 remaining observations were matched to 799 different electoral wards. A histogram of the observed IMD for these 799 wards is shown in Figure G.28 together with the equivalent histogram for all 8414 English wards. Higher values of the IMD indicate higher values of deprivation.

Figure G.28 Distribution of Index of Multiple Deprivation 2000 by electoral ward

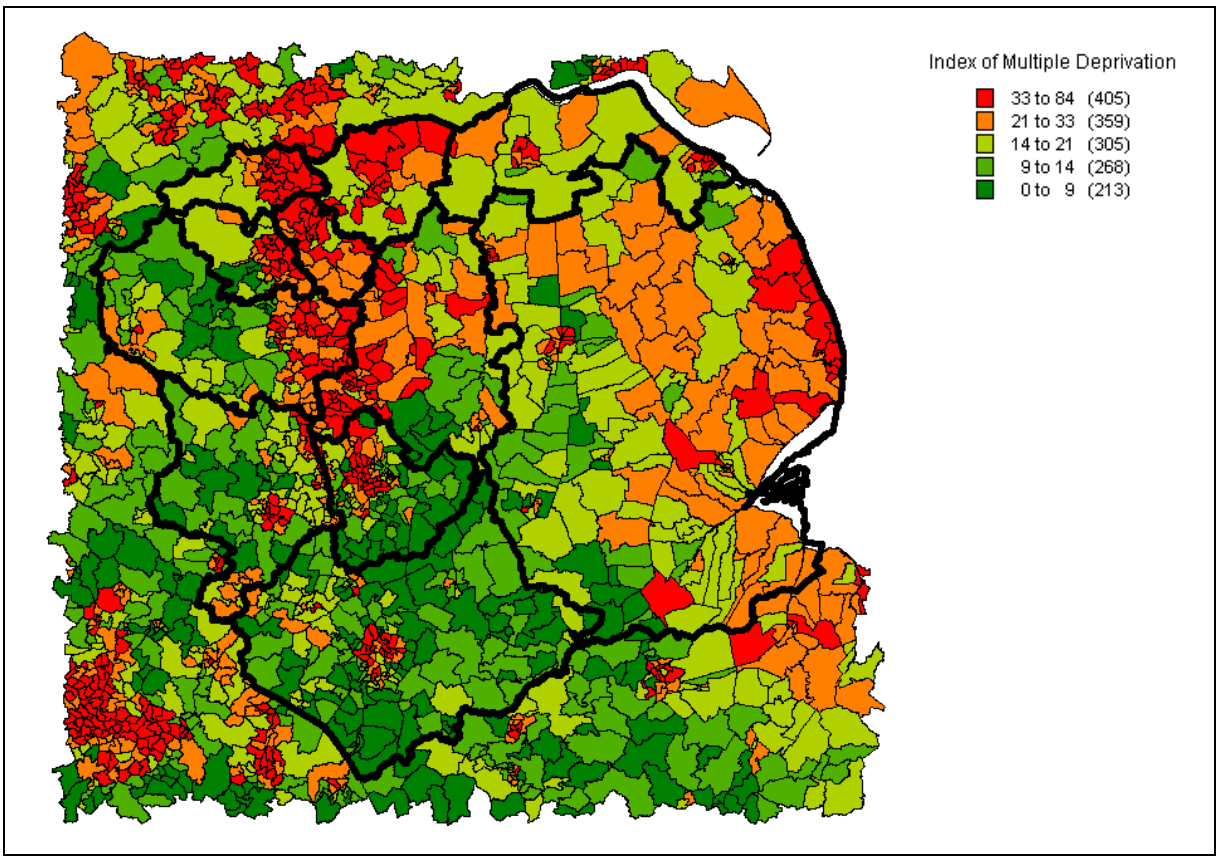


The minimum IMD score for the TNS data was 3.54 (ranked 8282 out of 8414 for all English wards) and the maximum was 73.48 (42 out of 8414 for all English wards). The median value for the 799 wards was 24.54, compared to 16.93 for England. These data, therefore, covered almost the full range of values seen in England, but the average value was higher.

Figure G.29 shows the value of the Index of Multiple Deprivation by electoral ward across the East Midlands. The values have been categorised into five groups according to the observed quintiles of the data for the whole of England. There were differences in deprivation, as measured by the IMD, between the different parts of the former Trent Health Authority area.

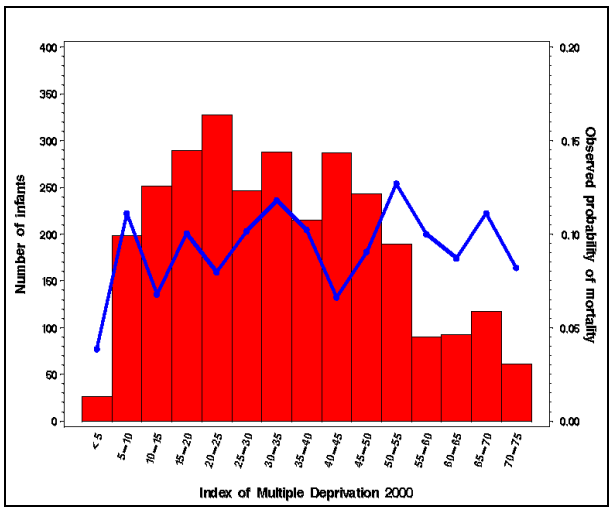
In particular, in the north and east of the Region areas of high deprivation predominated. However, a map such as Figure G.29 shows the score in relation to geographic area and not by population, and electoral wards tend to be small in area in high population urban areas and large in less densely populated rural areas.

Figure G.29 Map of Index of Multiple Deprivation 2000 by electoral ward



(This work is based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown.)

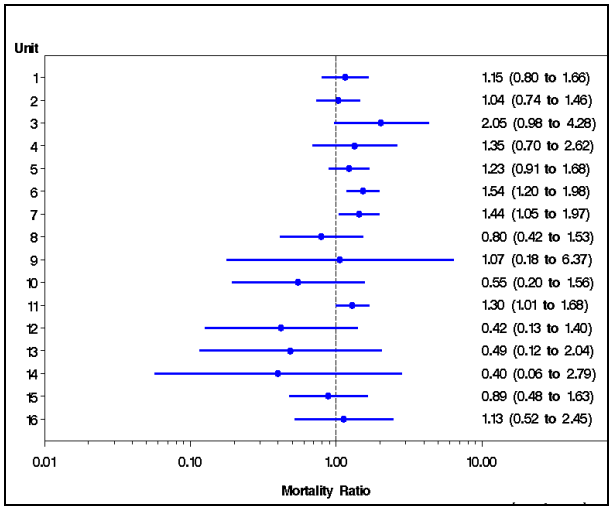
Figure G.30 Observed mortality by Index of Multiple Deprivation



The odds ratio for mortality for a unit increase in IMD was 1.00 (95% CI: 0.99 to 1.01), p = 0.49. When gestational age was included in the model (no evidence for an interaction by gestational age: p = 0.44) the odds ratio for a unit increase in IMD was 1.00 (95% CI: 0.99 to 1.01), p = 0.89 ($A_{ROC} = 0.886$: $\hat{C} = 5.55 \sim \chi^2_7$, p = 0.59).

There was no evidence that this relationship differed between units: p = 0.44. After adjustment for IMD and gestational age three units (6, 7 and 11) all had estimated 95% confidence interval wholly above the value one.

Figure G.31 Estimated standardized mortality ratios adjusted for Index of Multiple Deprivation and gestational age at birth



Appendix G.10 Antenatal corticosteroids

The use of antenatal corticosteroids prior to preterm birth has long been known to reduce subsequent respiratory distress syndrome (RDS) (Liggins and Howie, 1972) and, therefore, neonatal mortality (Crowley, 2003). Some 40-50% of all births under 33 weeks are affected by RDS (Chiswick, 1995). Guidelines from the Royal College of Obstetricians and Gynaecologists recommend that antenatal corticosteroids should be offered to women at risk of preterm delivery and that the optimal treatment-delivery interval is more than 24 hours but fewer than seven days after the start of treatment (Royal College of Obstetricians and Gynaecologists, 2004). Infants of mothers not given corticosteroids are more likely, therefore, to have poorer prognoses.

Table G.12 shows the proportion of infants whose mother received antenatal corticosteroids by hospital of birth (the codes used are the same as the codes used for the neonatal units

elsewhere in the thesis). There was strong statistical evidence that these rates differ ($p < 0.0001$) for the units investigated in this thesis; not including home births and those born outside of the former Trent Health Region. It would be of interest to investigate these differences to see whether they are the result of policy differences or due to differences in referral patterns or other clinical differences in the mothers or infants. However, this is not possible from TNS data and it will be merely reported that these differences still hold after the exclusion of in-utero transfers.

Table G.12 Antenatal corticosteroids by hospital of birth

Hospital of birth	No. Infants	No. antenatal corticosteroids	(%)
1	174	127	(73.0)
2	260	190	(73.1)
3	54	39	(72.2)
4	138	123	(89.1)
5	304	241	(79.3)
6	347	269	(77.5)
7	234	164	(70.1)
8	119	98	(82.4)
9	31	28	(90.3)
10	132	113	(85.6)
11	395	312	(79.0)
12	191	174	(91.1)
13	133	96	(72.2)
14	85	61	(71.8)
15	118	97	(82.2)
16	89	70	(78.7)
Home	25	1	(4.0)
Out of Region	196	94	(48.0)
Total	3025	2298	(76.0)

Obviously such differences in the administration of corticosteroids in the units of birth are likely to be reflected in the neonatal admissions. This can be seen in Table G.13. The proportion of infants whose mothers have received antenatal corticosteroids ranged from 63% (Unit 3) to over 91% (Unit 4).

Table G.13 Antenatal corticosteroids by NICU of care

Neonatal Unit	No. infants	No. antenatal corticosteroids	(%)
1	212	159	(75.0)
2	283	184	(62.0)
3	38	24	(63.2)
4	143	130	(90.9)
5	333	251	(75.4)
6	378	287	(75.9)
7	243	167	(68.7)
8	124	102	(82.3)
9	35	27	(77.1)
10	146	122	(83.6)
11	444	345	(77.7)
12	196	169	(86.2)
13	136	90	(66.2)
14	90	65	(72.2)
15	124	98	(79.0)
16	100	78	(78.0)
Total	**3025**	**2298**	**(76.0)**

Observed mortality by antenatal corticosteroid use is shown in Table G.14. Those infants who did not receive antenatal corticosteroids had a higher rate of mortality than those who did. This may have been due to the beneficial effect of corticosteroid administration or it may have been because those who did not receive it were extremely sick infants who needed to be delivered quickly. It is not possible to investigate this further with these data, nor is it possible to investigate the effect of the timing of corticosteroid use.

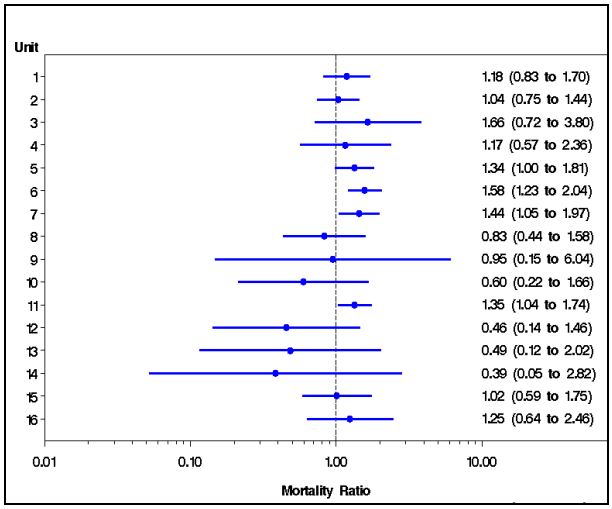
Table G.14 Mortality by antenatal corticosteroid use

Antenatal corticosteroids	No. Infants	No. died	(%)
Given	2298	192	(8.4)
Not given	727	93	(13.4)
Total	**3025**	**285**	**(9.4)**

The estimated odds ratio for mortality was 0.62 (95% CI: 0.47 to 0.81), p = 0.0004. When gestational age was included in the model (no evidence for an interaction by gestational age:

p = 0.46) the estimated odds ratio was 0.63 (95% CI: 0.45 to 0.87), p = 0.0047 (A_{ROC} = 0.883: $\hat{C}$ = 6.01 ~ χ^2_6, p = 0.42). There was no evidence that this relationship differed between units: p = 0.63. When SMRs were estimated, adjusted for antenatal corticosteroid administration and gestational age, Units 6, 7 and 11 had 95% confidence intervals completely above the value one.

Figure G.32 Estimated standardized mortality ratios adjusted for use of antenatal corticosteroids and gestational age at birth



Appendix G.11 Intrapartum monitoring

The term **fetal distress** is an often-used description of problems during birth, although it has no clear definition. In general, it describes the situation where the fetus is deprived of oxygen during labour or delivery, although this is often called acute fetal distress to distinguish it from sustained hypoxia during the pregnancy, which is usually termed chronic fetal distress. It has been suggested that a strict definition should include a combination of hypoxia, hypercarbia and acidosis (Mead, 1996). Such a reduction of oxygen in the blood (*hypoxia*) leads to an increase in carbon dioxide (*hypercarbia*), an increase in hydrogen ion concentration and, therefore, lower blood pH (*acidosis*). Such acidosis can result in increased neonatal morbidity (Winkler et al. 1991; Low et al. 1994; Nagel et al. 1995; Low et al. 1995a), although one of these studies has cast doubt on whether this still holds for infants born at less than 32 weeks gestational age (Low et al. 1995a). There are several potential causes of fetal distress (Table G.15).

Table G.15 Causes of fetal distress

Causes of fetal distress – from Beischer and Mackay (1978)	
Mother:	Hypotension or shock from any cause
	Cardiovascular disease
	Anaemia
	Respiratory depression or disease
	Malnutrition
	Acidosis and dehydration
Uterus:	Excessive or prolonged uterine activity
	Vascular degeneration
Placenta:	Premature separation
	Vascular degeneration and infarction
Cord:	Compression
Fetus:	Infection
	Malformation
	Haemorrhage
	Anaemia

Acute fetal distress may indicate an increased risk of perinatal mortality, cerebral palsy or neuro-developmental disability. A diagnosis of fetal distress is used to try to identify chronic fetal hypoxia to allow intervention before it leads to mortality or neurological damage. However, an abnormal fetal heart rate is not necessarily an indication of hypoxia but could indicate other pathologies, such as uterine rupture or fetal thyrotoxicosis (Royal College of Obstetricians and Gynaecologists, 2001:30). Indeed, it could be argued that the detection of fetal hypoxia, by EFM or Doppler, allows appropriate interventions to be carried out thus preventing, or at least limiting, any neurological disability. Such interventions include increasing oxygen to the mother, changing the mother's position, the use of tocolytic agents to relax the uterus, amnioinfusion and rapid delivery (Royal College of Obstetricians and Gynaecologists, 2001:58-61). However, there is no evidence from clinical trials to indicate the most appropriate method to manage births where fetal distress is suspected (Hofmeyr and Kulier, 2003). Good neonatal care can also often reverse the effects of acidosis (Eaton *et al,* 1994).

A number of different approaches have been advocated to try to detect fetal distress (Mead, 1996) and six variables are collected in TNS to try to identify those deliveries where fetal distress occurred: 'fetal distress', 'CTG abnormality', 'Doppler abnormality', 'abnormal scalp pH', meconium present' and 'other'. Each of these will be briefly investigated.

Fetal distress

The first variable records whether any signs of fetal distress during labour were noted, usually through the monitoring of the fetal heart rate, either from electronic fetal monitoring (EFM) or auscultation, or the monitoring of fetal movement. However, such observations may be poor indicators of fetal hypoxia and, therefore, of subsequent morbidity and mortality (Rosen and Dickinson, 1993; Parer, 2003). For the TNS data, there was no evidence for an association between mortality and these recorded signs of fetal distress (Table G.16): odds ratio = 0.99 (95% CI 0.76 to 1.28); p = 0.95.

Table G.16 Mortality by signs of fetal distress

EHM	No. infants	No. died	(%)
None	1948	184	(9.5)
Reported	1077	101	(9.4)
Total	3025	285	(9.4)

CTG abnormality

The next question on the TNS questionnaire records whether any abnormalities were noted from cardiotocogramography (CTG). This procedure looks in more detail at fetal heart function, with abnormalities falling into several different categories (Pearce and Steel, 1987:124-135). However, it is recognised that such fetal heart rate patterns can be difficult to interpret and such interpretation *"... is significantly affected by intra- and inter-observer error"* (Royal College of Obstetricians and Gynaecologists, 2001:51). However, there is some evidence of an association between fetal heart rate in the 24 hours before birth and neonatal mortality (Ayoubi *et al,* 2002). Information on the TNS form does not differentiate between the types of abnormality noted. There was no evidence for an association between mortality and CTG abnormality: odds ratio = 0.82 (0.62 to 1.09) p = 0.17.

Table G.17 Mortality by CTG abnormalities

CTG	No. infants	No. died	(%)
Normal	2155	213	(9.9)
Abnormal	870	72	(8.3)
Total	3025	285	(9.4)

Doppler abnormality

Ultrasound can be used to directly measure blood flow through the umbilical cord using Doppler frequency shift (Trudinger, 1999): intrapartum umbilical artery Doppler velocimetry. Although it has been suggested that Doppler velocimetry is a poor predictor of perinatal outcomes (Farrell *et al,* 1999), there was some evidence that abnormal patterns of blood flow are associated with increased neonatal mortality (Trudinger *et al,* 1991).

Table G.18 Mortality by abnormal Doppler velocimetry

Doppler	No. infants	No. died	(%)
Normal	2370	193	(8.1)
Abnormal	355	48	(13.5)
Missing	300	44	(14.7)
Total	3025	285	(9.4)

For those observations where Doppler velocimetry is known to have been carried out, there is strong evidence of an association between recorded abnormality and subsequent in-unit mortality: odds ratio = 1.76 (95% CI 1.25 to 2.48); p = 0.0009. This association still holds if the observations with abnormal Doppler are compared to all of the rest of the observations: odds ratio = 1.61 (95% CI 1.15 to 2.24) p = 0.0049.

Although these data indicate that abnormal umbilical artery blood flow was a predictor for mortality, the large proportion of infants did not undergo Doppler velocimetry (9.9%). Since the mortality rate for those without Doppler velocimetry was the highest of the three groups shown in Table G.18, it is unlikely that this group comprised solely those births thought to be uncomplicated. This makes the inclusion of Doppler velocimetry in a risk-adjustment model for all infants difficult.

Meconium present

A further observation recorded by TNS to try to identify fetal hypoxia is the presence of meconium (an infant's first faeces usually passed after birth) in the amniotic fluid. There is evidence in term infants of an association between meconium stained amniotic fluid and fetal hypoxia (Jazayeri *et al,* 2000), with meconium staining occurring in some 14% of all births (Ghidini and Spong, 2001).

Although hypoxia can be a cause of a fetus passing meconium into the amniotic fluid in-utero through stimulation of the vagus nerve, other risk factors can be involved (Miller *et al,* 1975), including difficult delivery, umbilical cord complications, poor intrauterine growth and other

chronic medical conditions. Meconium staining occurs predominately in term, or beyond, births and can be seen as an indicator of the increased maturation of the gastrointestinal tract. It is estimated that up to 30% of births beyond 42 weeks have meconium-staining (Creasy, 1997:111). The significance of meconium staining in very preterm births is less clear, with a recent small study showing a non-statistically significant increase in mortality for those infants with meconium staining: odds ratio 2.64 (95% CI 0.76 to 9.21) (Tybulewicz *et al,* 2004).

In addition to its association with hypoxia, the inhalation of the mixture of meconium and amniotic fluid by the infant can be a cause of Meconium Aspiration Syndrome (MAS) as the meconium traps air within the lungs and causes irritation to the airways. However, it is estimated that only 11% of infants with meconium staining develop MAS and even some of these cases may have causes other than the meconium inhalation (Ghidini and Spong, 2001).

The amount and type of meconium present in the amniotic fluid are important prognostic factors with fresh (green) meconium usually more associated with acute hypoxia than old (brown) meconium (Pearce and Steel, 1987:135). However, data recorded by TNS only indicates the presence of meconium. In these data there was no evidence of an association between the presence of meconium-staining of the amniotic fluid and subsequent in-unit mortality: odds ratio = 1.06 (95% CI 0.54 to 2.06); p = 0.87.

Table G.19 Mortality by presence of meconium

Meconium-staining	No. infants	No. died	(%)
None	2924	275	(9.4)
Present	101	10	(9.9)
Total	3025	285	(9.4)

Other indicators of fetal distress

Other significant events that may indicate fetal distress (e.g. cord prolapse, fetal distress in the other twin) are recorded by a separate question on the TNS form. There was no evidence of an association between these other indicators of fetal distress and subsequent in-unit mortality (Table G.20): odds ratio = 0.99 (95% CI 0.76 to 1.28) p = 0.95.

Table G.20 Mortality by other indicator of intrapartum difficulties

Other	No. infants	No. died	(%)
None	1948	184	(9.5)
Reported	1077	101	(9.4)
Total	3025	285	(9.4)

Abnormal scalp pH

The final indicator of fetal distress recorded by TNS is scalp blood pH. This is a direct measure of acidosis and can be performed at an early stage of labour. A sample of blood is taken directly from the fetus's scalp when the fetus is in cephalic presentation, or the buttock if not (Al-Azzawi, 1990).

A study using sheep, sampling scalp blood while simultaneously directly measuring the pH of the fetal preductal arterial blood, has shown evidence that the two sources of measurement correlated well (Morgan *et al,* 2002). For TNS pH less than 7.25 was considered abnormal for all births, as there is evidence that blood gas levels are the same in otherwise uncomplicated preterm births as in term births (Ramin *et al,* 1989).

Only five infants were recorded as having abnormal scalp pH of less that 7.25 (Table G.21). There is little statistical power to investigate an association between abnormal scalp pH and subsequent in-unit mortality: odds ratio for normal vs. abnormal = 2.45 (95% CI 0.27 to 22.05) p = 0.41.

Table G.21 Mortality by scalp pH

Scalp pH	No. infants	No. died	(%)
Normal (≥ 7.25)	2379	220	(9.3)
Abnormal (< 7.25)	5	1	(20.0)
Missing	641	64	(10.0)
Total	3025	285	(9.4)

As scalp pH is a direct measure of blood gasses it is of interest to see how well the indirect measures of fetal distress correlate with abnormal scalp pH (Table G.22).

Table G.22 Scalp pH by other measures of fetal distress

pH	Distress		CTG		Doppler		Meconium	
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Normal	1476	903	1677	702	2098	281	2306	73
Abnormal	3	2	2	3	5	0	5	0
Missing	469	172	476	165	567	74	613	28

Even allowing for the small number of infants with abnormal scalp pH, that there is little correlation with the different methods. Indeed, for abnormal Doppler velocimetry and meconium staining no infant with recorded abnormality had abnormal scalp pH. This may indicate that these indirect measures are of little use. However, an association between abnormal umbilical artery blood flow and mortality has been demonstrated in these data.

Combined indicator

It has been suggested previously that such indirect measures that try to indicate hypoxia are poor predictors, generating a very high proportion of false positives (Low *et al,* 1995b). The combined use of indicators may be of more use, for example performing fetal blood analysis when the CTG is abnormal (Saling, 1996). As a final approach, an infant with any of the indicators recorded as abnormal was assumed to have experienced fetal distress (Table G.23).

Table G.23 Mortality by fetal distress

Fetal distress	No. infants	No. died	(%)
None	1822	178	(9.8)
Reported	1203	107	(8.9)
Total	3025	285	(9.4)

The estimated odds ratio for mortality was 0.90 (95% CI: 0.70 to 1.16), p = 0.42. When gestational age at birth was included in the model there was evidence for a fetal distress by gestational age interaction (p = 0.016).

$$\hat{g}_i = \hat{\beta}_0 + \hat{\beta}_D . distress_i + \hat{\beta}_G . gest_i + \hat{\beta}_{DG} . distress_i . gest_i$$

$\hat{\beta}_0 = 18.44$ (s.e. 1.25)

$\hat{\beta}_D = -3.98$ (s.e. 1.91)

$\hat{\beta}_G = -0.76$ (s.e. 0.05)

$\hat{\beta}_{DG} = 0.17$ (s.e. 0.07)

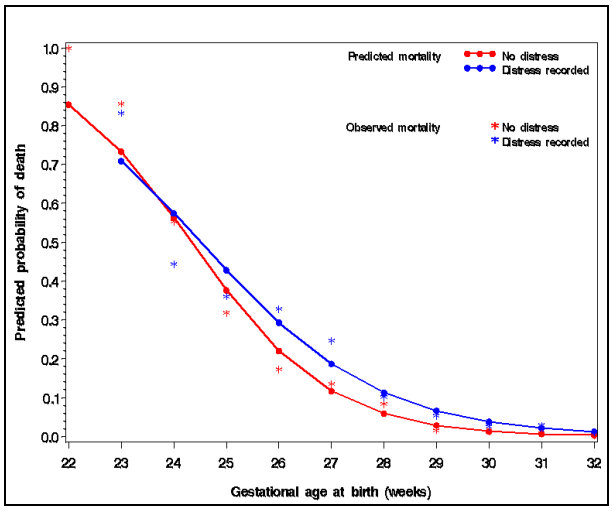
$$(A_{ROC} = 0.886:\ \hat{C} = 4.80\ \sim \chi^2_7,\ p = 0.68)$$

where: $distress = \begin{cases} 1 & \text{if sign of fetal distress recorded} \\ 0 & \text{otherwise} \end{cases}$

gest = gestational age at birth in completed weeks

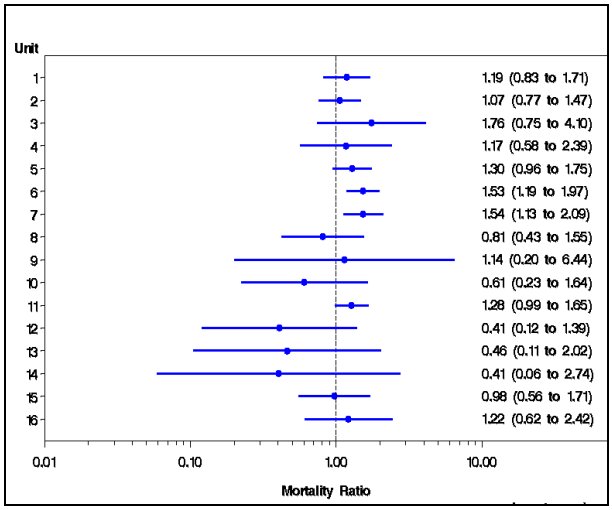
There was no evidence that the relationship between reported fetal distress and mortality differed amongst the units (p = 0.96). The estimated functions suggest that fetal distress was associated with mortality for infants born at 25 to 30 weeks gestational age (Figure G.33).

Figure G.33 Estimated mortality by recorded fetal distress and gestational age at birth



When SMRs were estimated (Figure G.34), adjusted for fetal distress and gestational age, Units 6 and 7 had 95% confidence intervals wholly above unity.

Figure G.34 Estimated standardized mortality ratios adjusted for fetal distress and gestational age at birth



Appendix G.12 Mode of delivery

The reported mode of delivery is shown in Table G.24. Twenty-five observations had unknown modes of delivery, none of whom died before discharge.

Table G.24 Mode of delivery

Mode of delivery		No. infants	(%)
Vaginal	**Normal**	1031	34.1
	Low forceps	42	1.4
	High forceps	25	0.8
	Ventouse	4	0.1
	Assisted breech	217	7.2
Caesarean section	**Emergency: Labouring**	480	15.9
	Emergency: not labouring	1122	37.1
	Elective	79	2.6
Missing		25	0.8

There is conflicting evidence on whether caesarean section reduces mortality for preterm infants (Penn and Ghaem-Maghami, 2001). Observational studies using specific populations have been equivocal, with some evidence presented for increased survival following caesarean section (Naylor *et al,* 2001), one study showed no evidence for a difference in long-term survival between vaginally and caesarean section delivered infants (Wolf *et al,* 1999) and a further study reported evidence for increased survival following vaginal delivery (Bauer *et al,* 2003). A Cochrane review of clinical trials found no evidence for an increase in neonatal survival with elective labouring caesarean section when compared to a policy of caesarean sections only if a clear clinical indication arose (Grant and Glazener, 2003). However, this review comprised only three small studies, as such studies encounter difficulties in recruitment (Penn and Steer, 1990). Guidelines from the National Institute for Clinical Excellence (NICE) state that, because of the uncertainty over the outcomes from planned caesarean section for preterm births, planned caesarean sections *"should not be routinely offered outside a research context"* (National Institute for Clinical Excellence, 2004:1.2.3.1). Seventy-nine infants from the TNS data were admitted following elective caesarean section (Table G.24) and these occurred at obstetric units in ten different hospitals.

For term births at least, labour may cause an increase in the white blood cell count of the infant (neutrophil leukocytosis) by delaying apoptosis and by increasing lipopolysaccharide (LPS) responsiveness (Molloy *et al,* 2004). Such changes increase the immunological ability of the infant, thus decreasing the risk of infection and mortality. It is not known whether labour promotes such physiological changes in preterm births, but it is known that premature infants have decreased neutrphil production compared to term infants (Carr, 2000), thus making potential labour-induced neutrophil leukocytosis especially important.

In order to obtain reasonably sized groups, the methods of delivery recorded by TNS were combined into three clinically homogeneous groups. In the light of the potential differences discussed above, the three categories were vaginal delivery (n = 1319, 44.0%), labouring caesarean section (n = 480, 16.0%) and non-labouring caesarean section (n = 1201, 40%) (Field, D.J.: Personal communication).

Infants delivered by caesarean section had statistically significant higher rates of survival: overall p-value = 0.0043 (Table G.25).

Table G.25 Unadjusted odds ratio for mortality by mode of delivery

Mode of delivery	No. infants	No. died	(%)	Odds ratio	(95% CI)	p-value
Vaginal	1319	151	11.5	reference		
CS: labouring	480	34	7.1	0.59	(0.40 to 0.87)	0.0076
CS: non-labouring	1201	100	8.3	0.70	(0.53 to 0.92)	0.0092
Total	3000	285	9.5			

However, the relative frequency of the modes of delivery changes with gestational age. At very early gestational ages most births are vaginal deliveries. As gestational age at birth increases, more infants are born by elective or non-labouring caesarean section (Figure G.35). Since mortality rates are much higher for the very earliest deliveries, gestational age may act as an effect modifier. Once gestational age was included in the model (p-value for interaction = 0.25) the situation was reversed: infants born by vaginal delivery had the lowest gestational age specific mortality rates (Table G.26).

Figure G.35 Mode of delivery by gestational age at birth

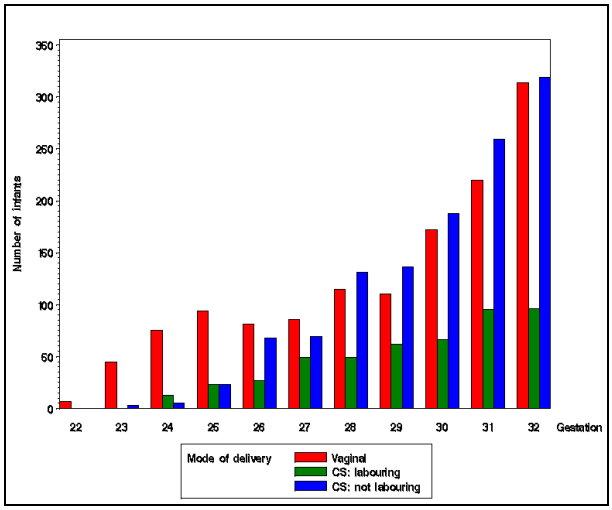
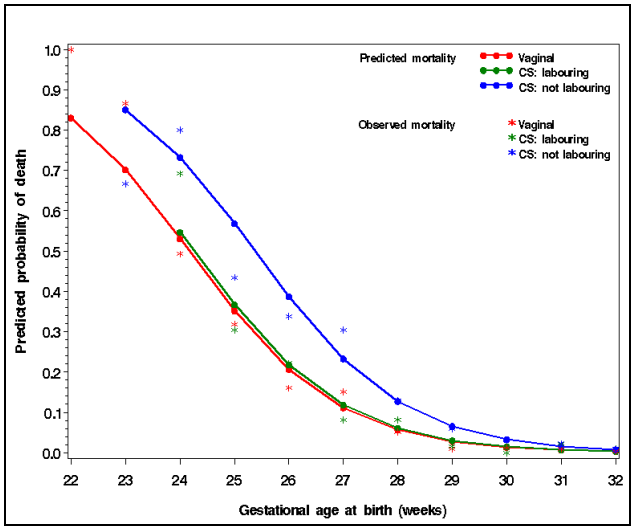


Table G.26 Odds ratio for mortality by mode of delivery adjusted for gestational age

Mode of delivery	Odds ratio	(95% CI)	p-value
Vaginal delivery	reference		
CS: labouring	1.07	(0.67 to 1.68)	0.78
CS: not labouring	2.42	(1.70 to 3.44)	< 0.0001
($A_{ROC} = 0.890$: $\hat{C} = 4.57 \sim \chi^2_8$, p = 0.80)			

This produced an interesting split between those deliveries where labour occurred (vaginal and labouring caesarean section) and those where it did not (both non-labouring and elective caesarean section).

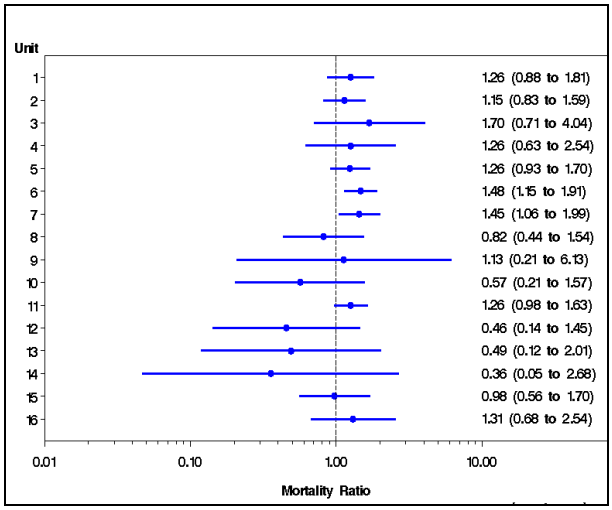
Figure G.36 Estimated mortality by mode of delivery and gestational age at birth



This may be because caesarean section without labour indicates that the fetus was having severe problems: there has been evidence presented that electively delivered preterm infants tend to be lighter than those from spontaneously labouring deliveries (Yudkin *et al,* 1987). On the other hand, the difference in outcomes may be because labour causes physiological changes in the fetus to its benefit. However, the TNS data do not allow such an investigation.

There was no evidence for an interaction between mode of delivery and NICU: p = 0.98. Standardized for gestational age and mode of delivery, Units 6 and 7 had 95% confidence intervals for the SMR wholly greater than unity.

Figure G.37 Estimated standardized mortality ratios adjusted for mode of delivery and gestational age at birth



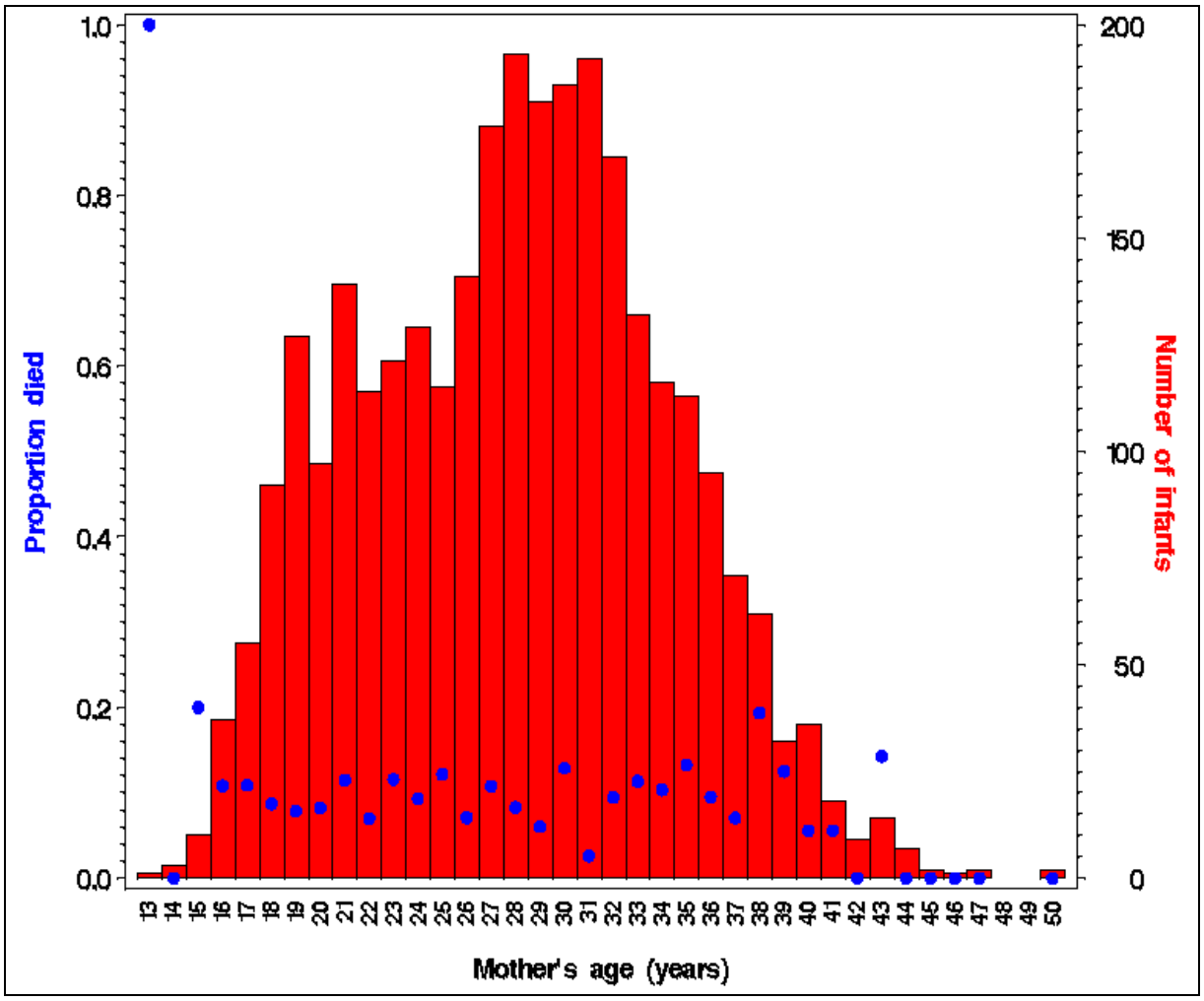
Appendix G.13 Mother's age

There is a large body of evidence that the risk of complications during pregnancy, and at delivery, increases with increasing maternal age (Fretts *et al,* 1995; Jolly *et al,* 2000; Temmerman *et al,* 2004). However, the link between neonatal outcomes and maternal age is less clear, with some evidence that there is no increased risk of poor neonatal outcomes with increased maternal age (Berkowitz *et al,* 1990).

For the TNS data, the observed age of the mothers ranged from 13 to 50 years (mean and median 28.0 years). There were 34 missing observations, of which 4 (11.8%) of the infants died before discharge. Inspection of the observed mortality by maternal age shows no evidence for increased mortality with greater age (Figure G.38). Indeed the very oldest

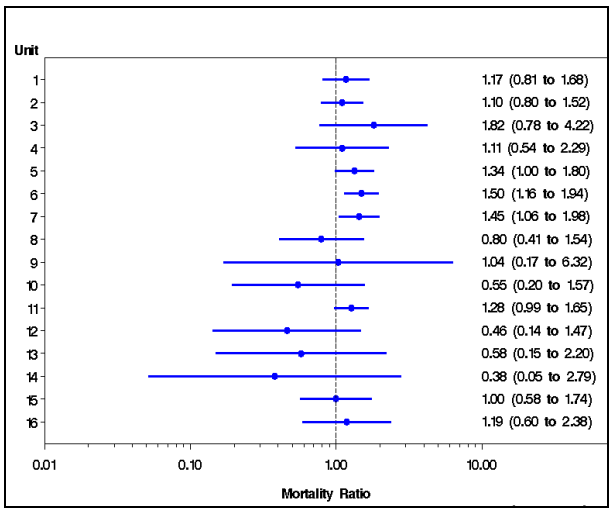
mothers (>40 years) appear to have the lowest rates of infant mortality, although the number of observations is quite small.

Figure G.38 Mortality by mother's age



Using a logistic regression model, there was no evidence for a relationship between a mother's age and the infant's risk of death: estimated odds ratio = 1.00; 95% CI 0.98 to 1.02; p = 0.98. This did not change after including gestational age at birth in the model: estimated odds ratio = 0.99; 95% CI 0.97 to 1.02; p = 0.54 (A_{ROC} = 0.884: $\hat{C}$ = 10.48 ~ χ^2_8, p = 0.23). There was also no evidence that the relationship between the mothers' ages and infant mortality varied in the different neonatal units: p = 0.25. After adjustment Units 6 and 7 had 95% confidence intervals for the SMR wholly above the value one (Figure G.39).

Figure G.39 Estimated standardized mortality ratios adjusted for mother's age and gestational age at birth



Appendix G.14 Previous obstetric history

A mother's previous obstetric history is usually quantified using **gravidity** and **parity**. Gravidity is defined as the number of previous pregnancies experienced by a woman. Parity is defined either as the total number of previous live births and stillbirths or, for hospital in-patient statistics, the total number of pregnancies leading to at least one registerable birth (i.e. $\geq$ 24 weeks) (Macfarlane and Mugford, 2000:13-14). The relationship between a mother's obstetric history and the probability of mortality for a subsequent infant is unclear. There has some evidence presented for a U-shaped relationship between neonatal mortality and parity (Bai et al. 2002), seen in Figure G.40.

Figure G.40 Neonatal mortality by parity (from Bai 2002)

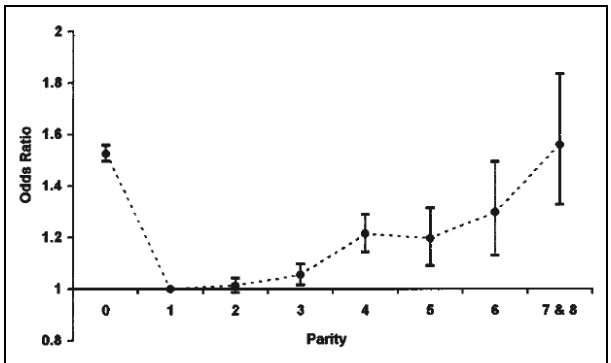
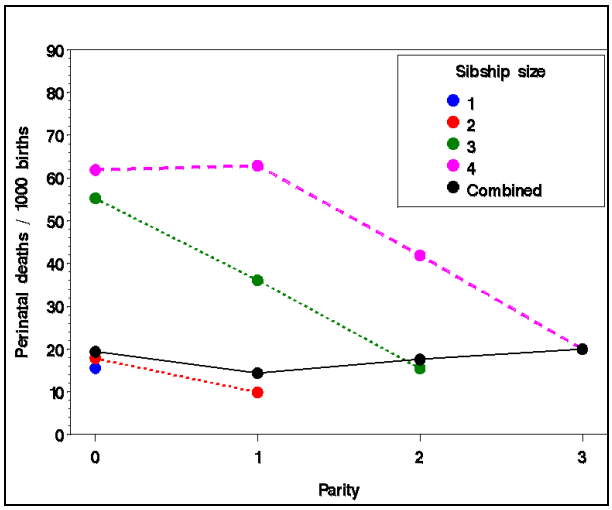


Fig 2. The risk of neonatal morbidity for different parity groups after age, smoking, English-speaking background, and private health insurance status have been taken into account. Parity = 1, as a reference group. *Error bars* represent 95% CI.

However, Bai *et al* used a cross-sectional design for their study and this approach may be unsuitable. Longitudinal study designs have reported decreasing rates of poor outcome with increased gravidity or parity (Billewicz, 1973; Roman *et al,* 1978; Bakketeig and Hoffman, 1979). Figure G.41 is produced using data from Bakketeig & Hoffman (1979) who studied all fetal deaths and live births at 16 weeks or more gestational age in Norway over the seven-year period 1967 to 1973. They were able to link births to the same mother in order to investigate the longitudinal relationship between previous pregnancies and subsequent outcomes. The data shown in Figure G.41 show outcomes to mothers who had no pregnancies before 1967. If the whole cohort is considered (black line) there is evidence for a U-shaped relationship similar to that in Figure G.40. However, if the data are grouped according to ultimate sibship size there is evidence of decreasing perinatal mortality with increasing parity.

Figure G.41 Perinatal mortality by parity (data from Bakketeig 1979)



This decrease in mortality could be due to 'self-selection', with mothers tending to become pregnant again after a pregnancy with an adverse outcome and, on the other hand, a higher probability of stopping childbearing after a successful pregnancy (Bakketeig and Hoffman, 1979). Thus, it is important to know the outcomes of previous pregnancies, not just the number (Yudkin, 1980). However, the true processes are not known.

In this thesis, for simplicity, only the previous number of pregnancies was considered but it is acknowledged that this has limitations. To avoid small numbers of observations in each group, the observations were divided into three categories: primagravida, secundigravida, multigravida.

Table G.27 Observed mortality by gravidity

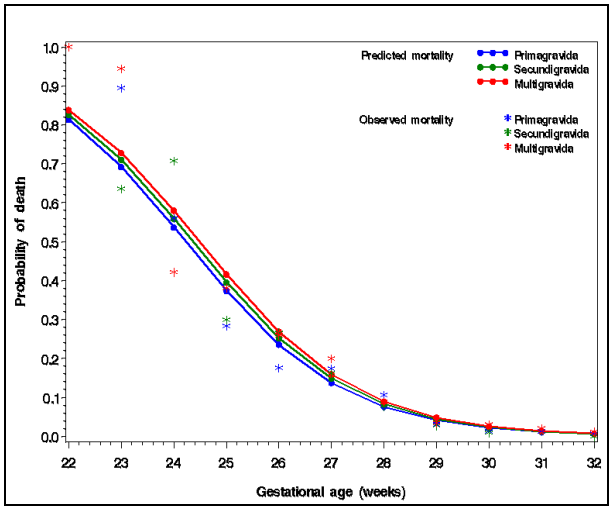
Gravidity	Total	Died	(%)	Odds ratio	(95% CI)	p-value
1	1163	93	(8.0)	reference		
2	737	66	(9.0)	1.13	(0.81 to 1.58)	0.46
3+	1125	126	(11.2)	1.45	(1.09 to 3.93)	0.0095

There was statistical evidence for difference in the mortality rates between the groups: p = 0.029 (Table G.27). However, once gestational age is included in the model (p-value for interaction = 0.28), there was no evidence for a difference between the groups in gestational age specific mortality rates: p = 0.58. The predicted mortality rates are shown in Figure G.42. There was no evidence that the relationship between the groups and gestational age differed by unit: p = 0.74.

Table G.28 Odds ratio for mortality by gravidity adjusted for gestational age

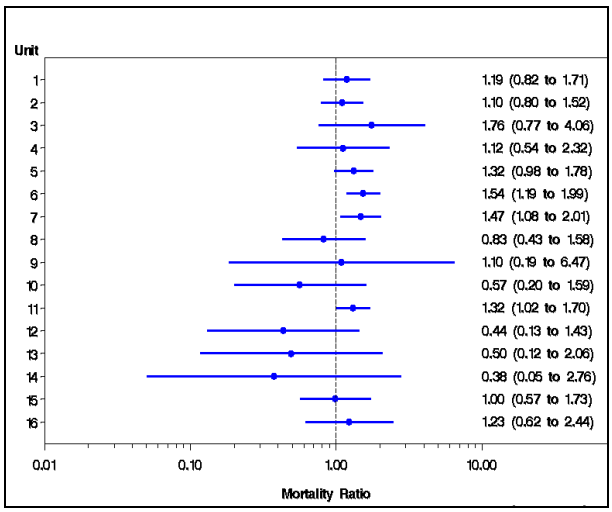
Gravidity	Odds ratio	(95% CI)	p-value
1	reference		
2	1.09	(0.74 to 1.62)	0.65
3+	1.19	(0.85 to 1.67)	0.30
(A_{ROC} = 0.883: $\hat{C}$ = 5.85 ~ χ^2_8, p = 0.75)			

Figure G.42 Mortality by gravidity and gestational age at birth



After adjustment Units 6, 7 and 11 had 95% confidence intervals for the SMR wholly above unity (Figure G.43).

Figure G.43 Estimated standardized mortality ratios adjusted for gravidity and gestational age at birth



Appendix G.15 Maternal or fetal infection

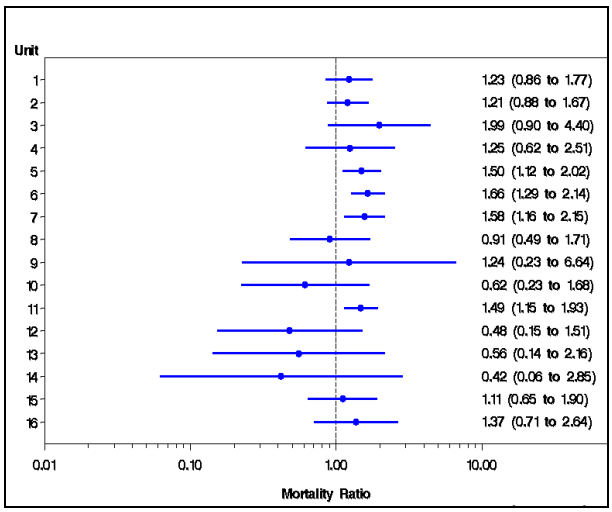
Maternal or fetal infection is known to increase the risk of preterm birth and to increase mortality among preterm infants (Fung *et al,* 2003; Garite and Freeman, 1982; Ernest, 1998). Many such infections are the result of preterm rupture of the membranes (Romero *et al,* 2003) and there is evidence that very preterm infants (<28 weeks gestational age) are at particularly increased risk of infection and mortality (Nelson *et al,* 1994).

Table G.29 Mortality by maternal or fetal infection

EHM	No. infants	No. died	(%)
None	2457	223	(9.1)
Infection	568	62	(10.9)
Total	3025	285	(9.4)

The observed odds ratio for mortality was 1.23 (95% CI: 0.91 to 1.66), p = 0.18. After adjustment for gestational age (p = 0.72 for interaction between infection and gestational age) the adjusted odds ratio was 0.84 (95% CI: 0.59 to 1.21), p = 0.34 (A_{ROC} = 0.882: $\hat{C}$ = 4.11 ~ χ^2_7, p = 0.77). There was no evidence that the relationship between infection and gestational age differed by unit: p = 0.95. After adjustment for infection and gestational age four units had confidence intervals wholly above the value one: Units 5, 6, 7 and 11 (Figure G.44).

Figure G.44 Estimated standardized mortality ratios adjusted for infection and gestational age at birth

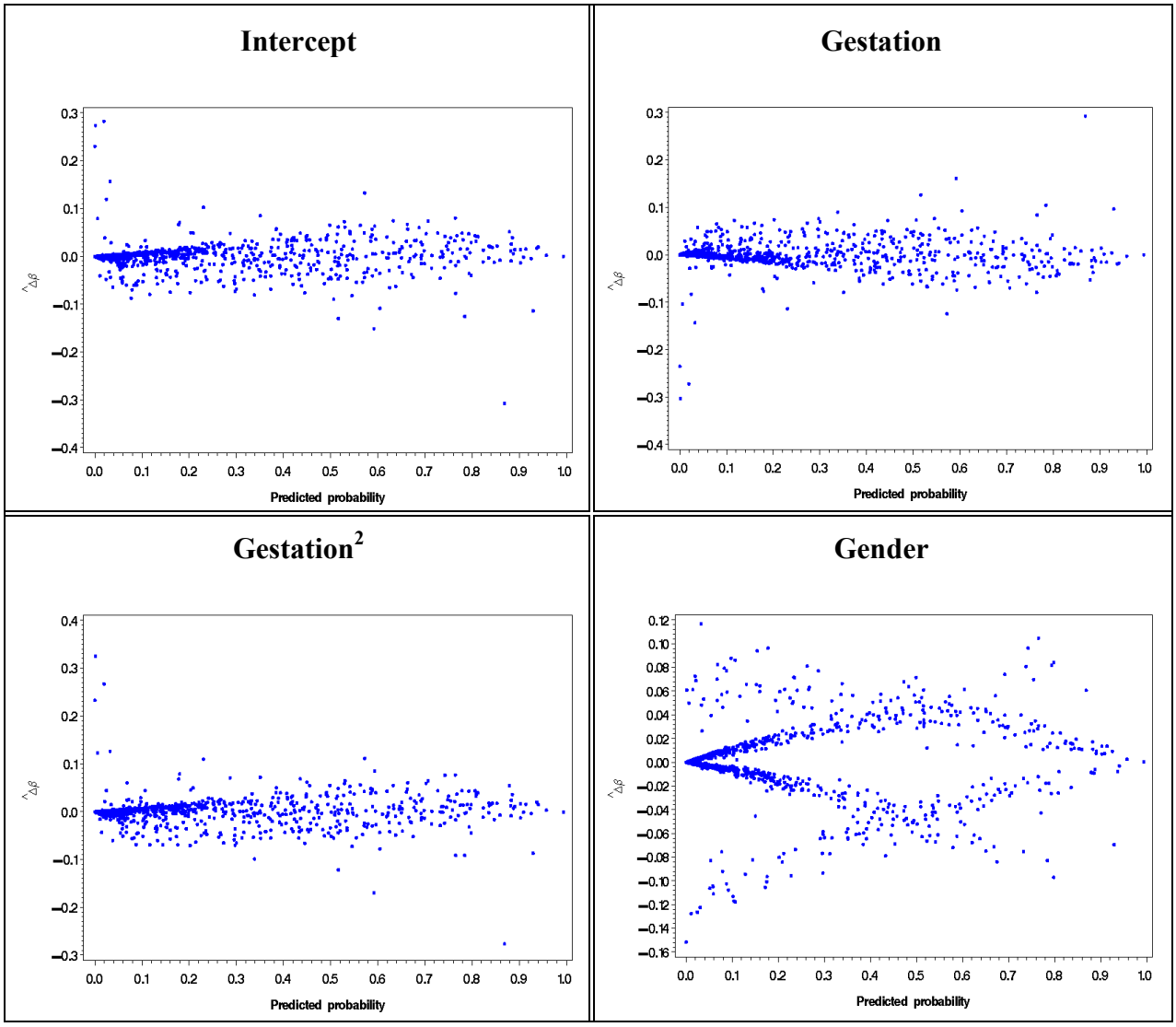


However, maternal or fetal infection, as recorded by TNS, is identified by the use of antibiotics in the mothers. This definition was originally selected with the aim of identifying those with strong evidence of infection (Field, D.J.: Personal communication) but it is recognised that the use of treatment as the indicator to detect infection raises the clear possibility of differential measurement error across the neonatal units. However, the TNS data do not allow an investigation into this question.

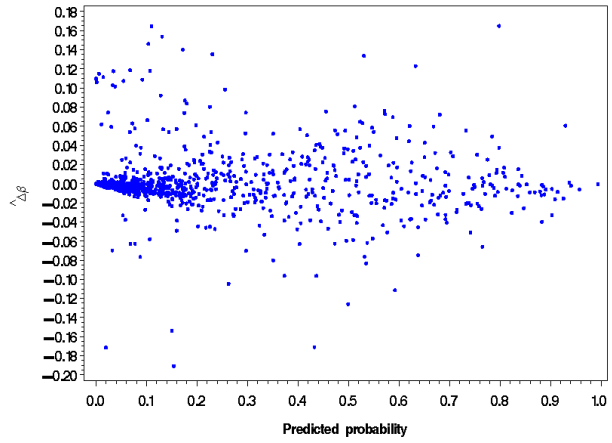
Appendix H: DF BETAs

In this appendix the DF BETAs from the 'Full' and 'Reduced' models are shown.

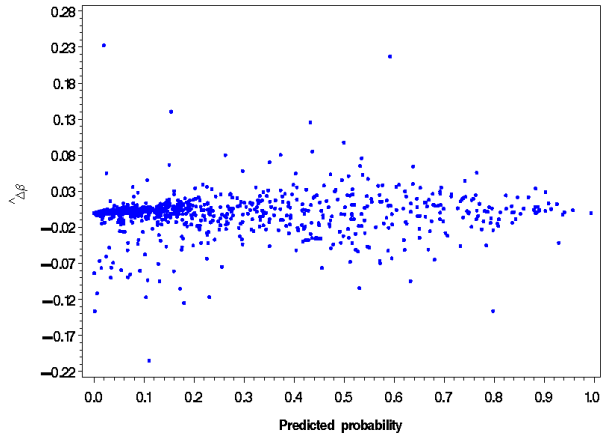
Figure H.1 DFBETAs for Full Model



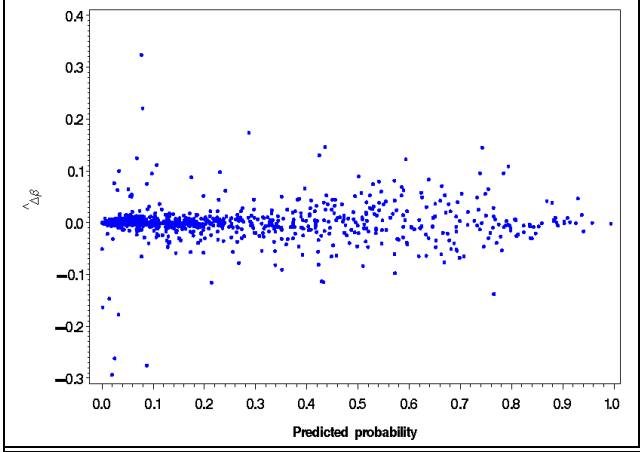
Birth weight



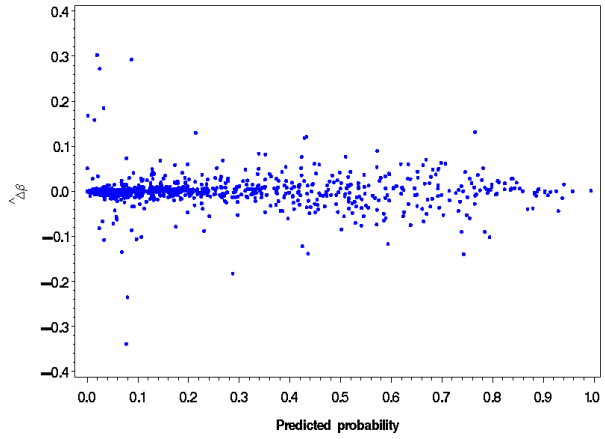
(Birth weight)^2



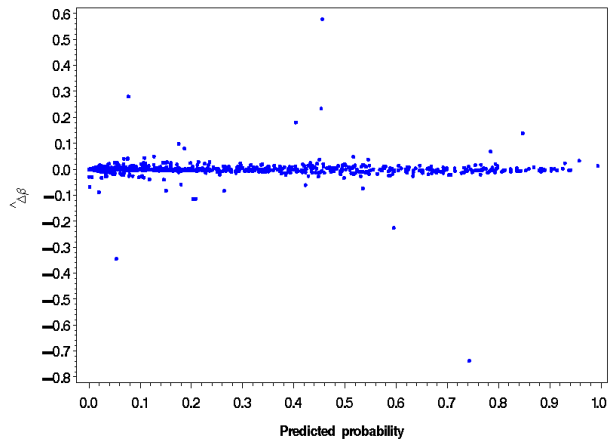
Apgar



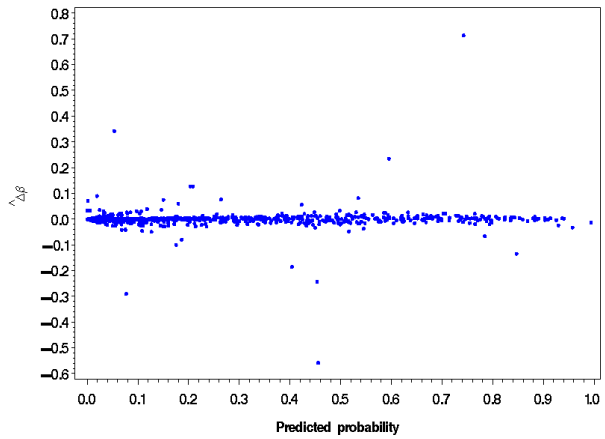
Apgar*Gestation



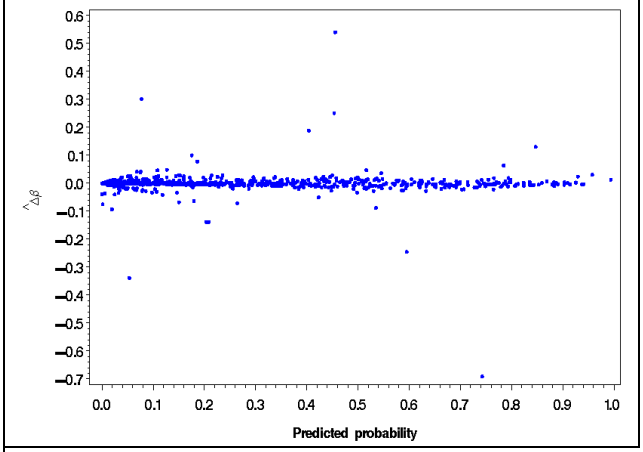
Congenital malformation



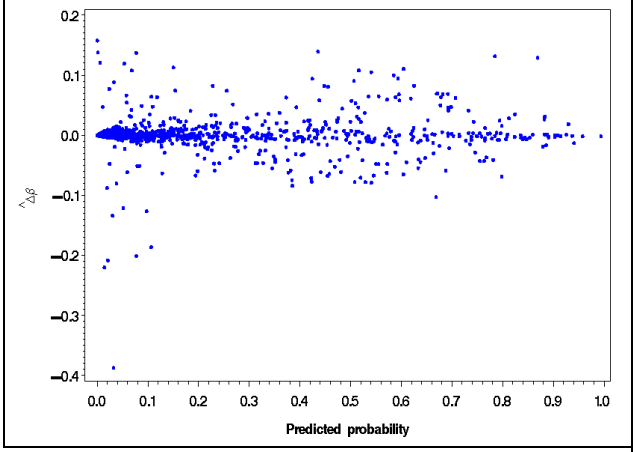
Congenital malformation*Gestational age



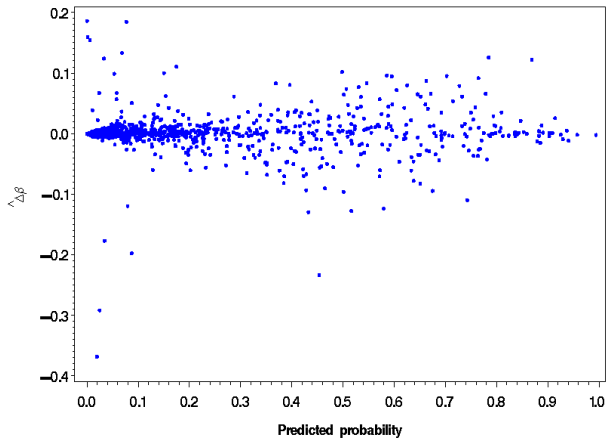
Congenital malformation*Gestational age^2



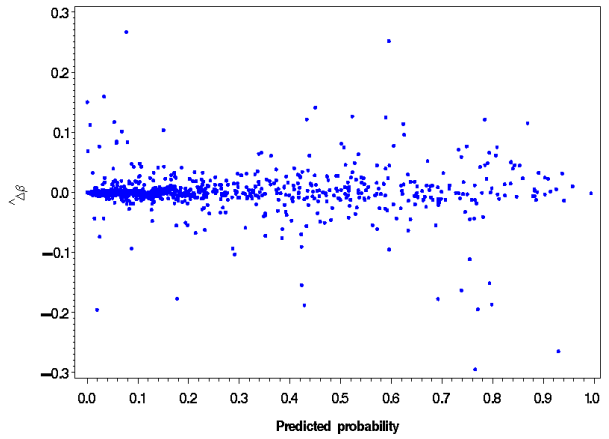
Base excess -7.0 to –9.9



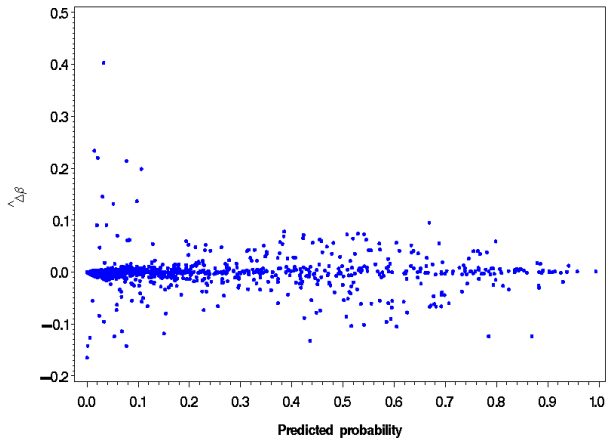
Base excess -10.0 to –14.9



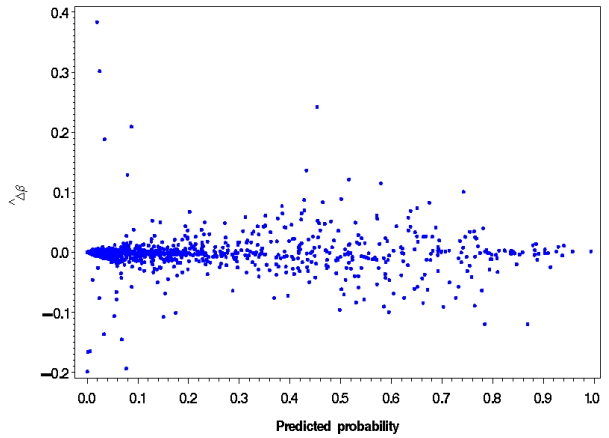
Base excess <-15



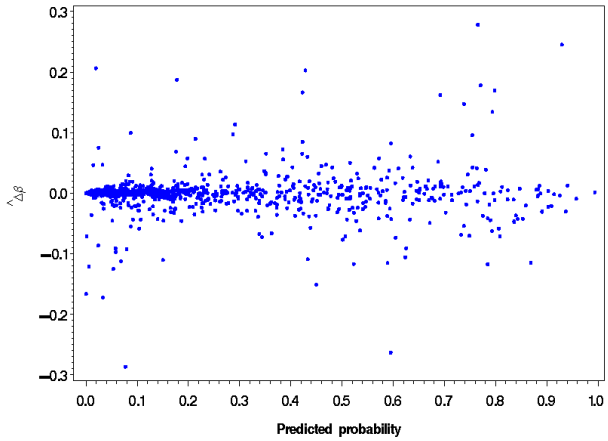
Base excess -7.0 to –9.9*Gestation



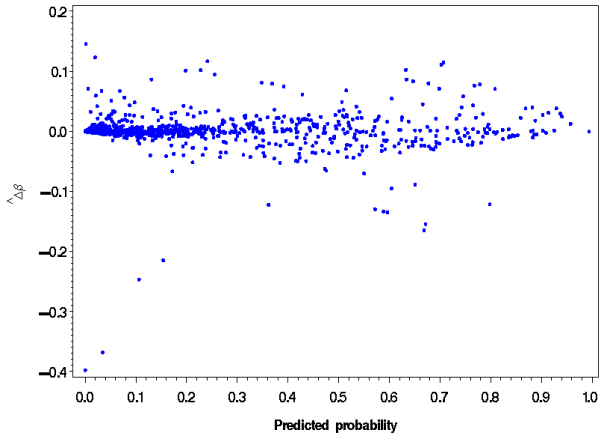
Base excess -10.0 to –14.9* Gestation



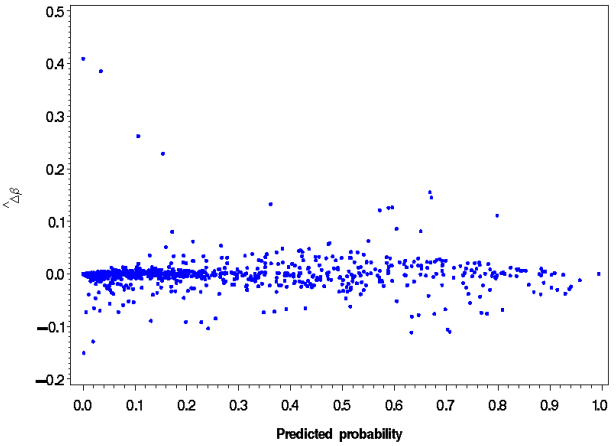
Base excess <-15* Gestation



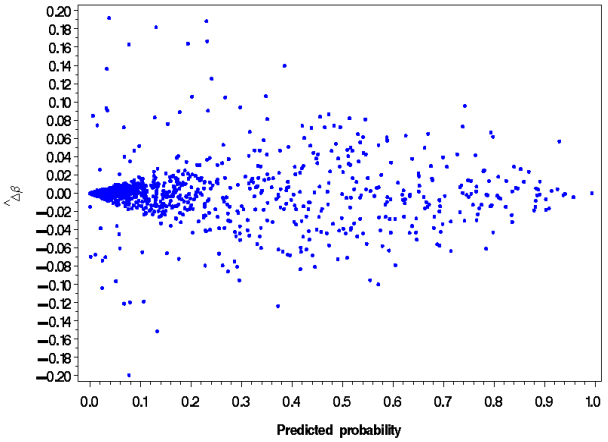
Multiple birth



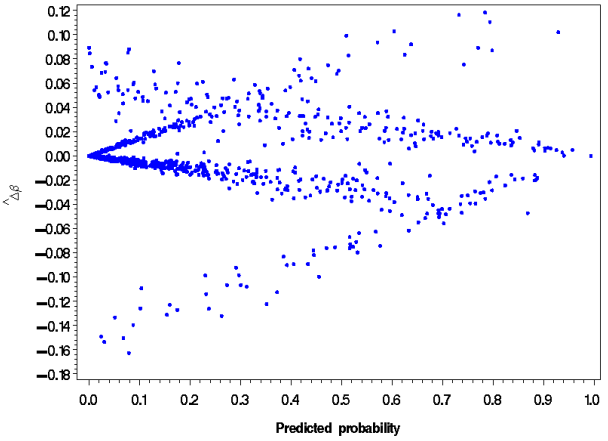
Multiple birth*Gestational age



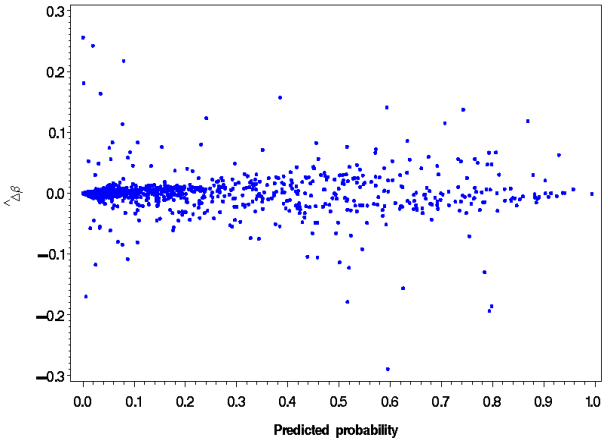
IMD



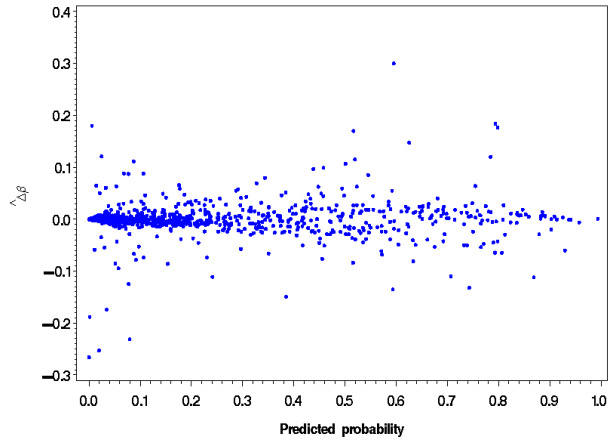
Corticosteroids



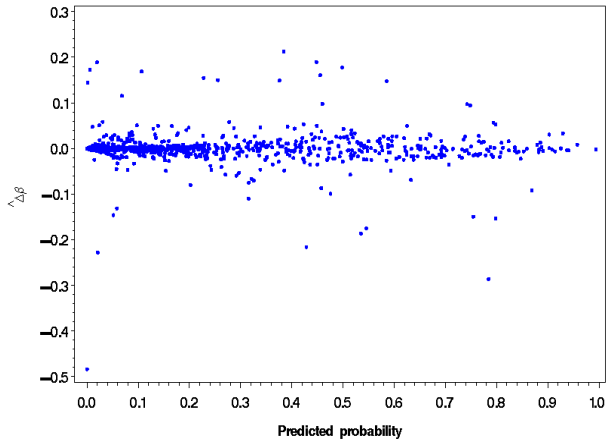
Fetal distress



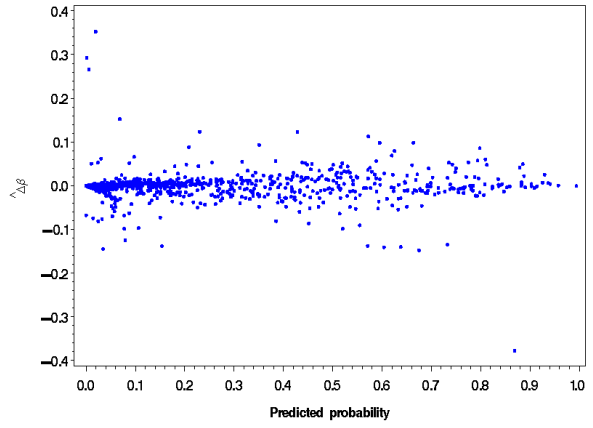
Fetal distress* Gestation



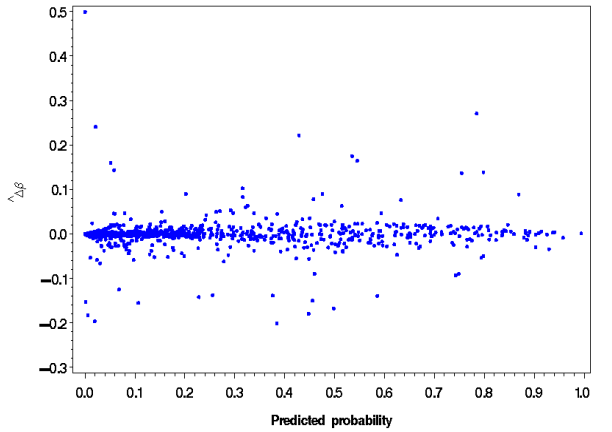
Mode of delivery: CS labouring



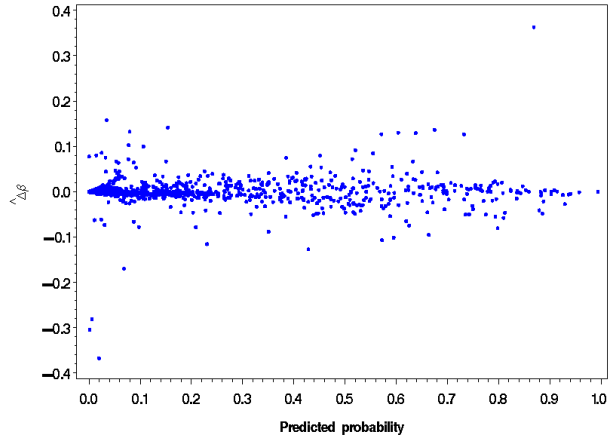
Mode of delivery: CS non-labouring



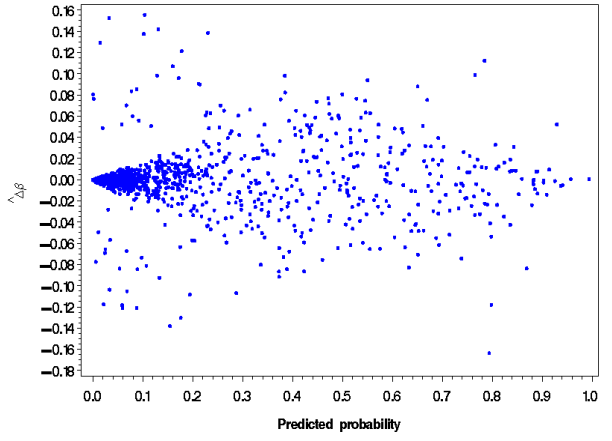
Delivery: CS labouring*Gestation



Delivery: CS non-labouring*Gestation



Mother's age



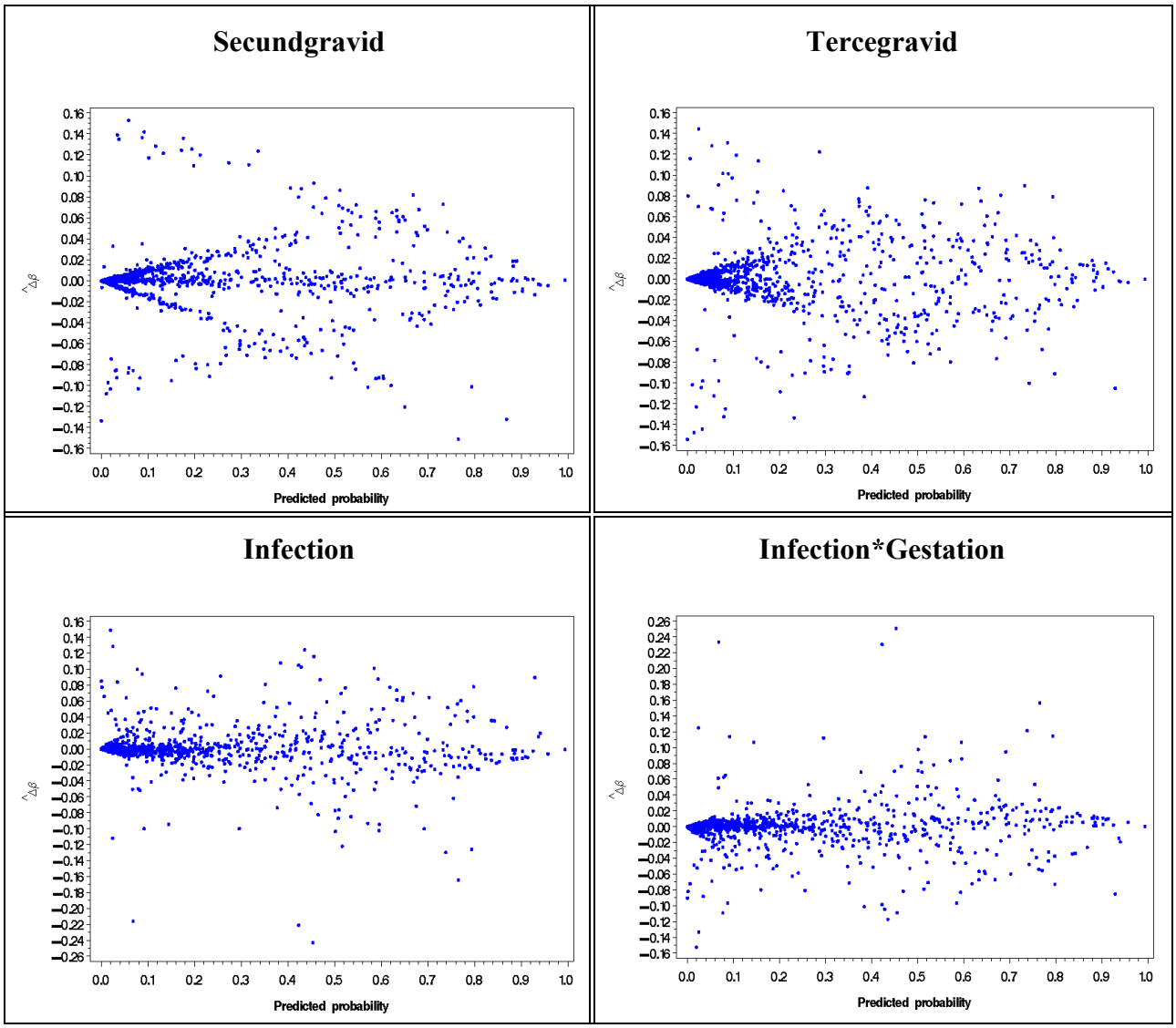
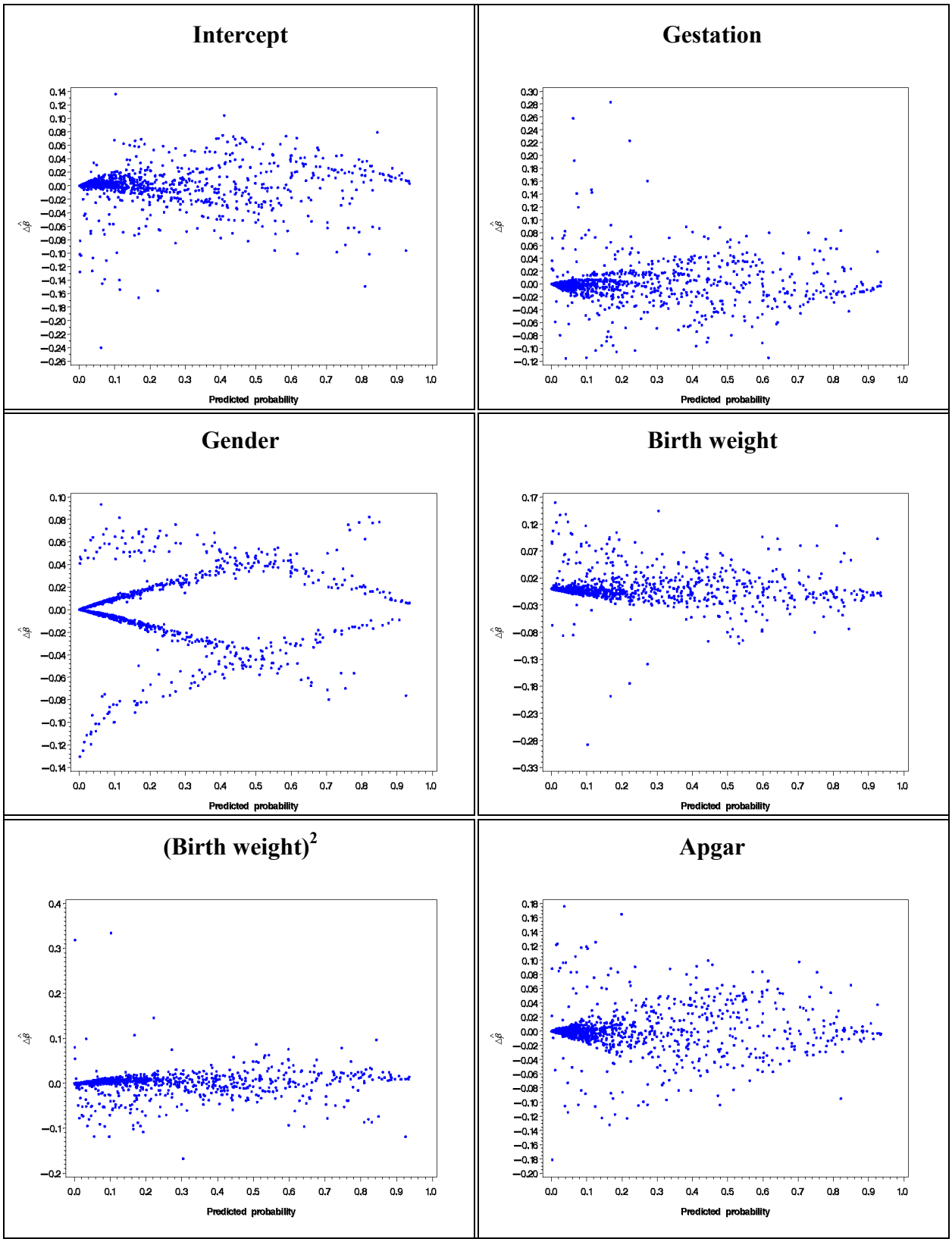
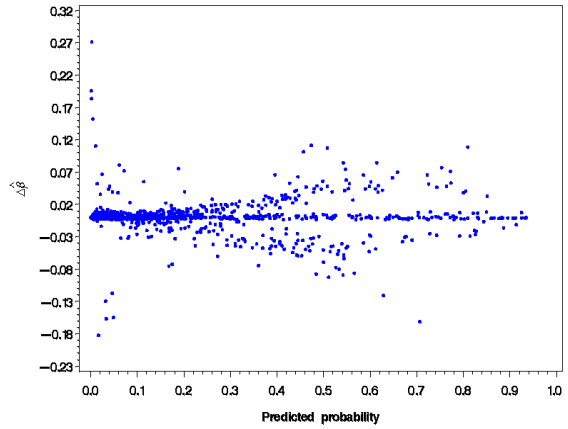

Secundgravid
Tercegravid
Infection
Infection*Gestation
Predicted probability

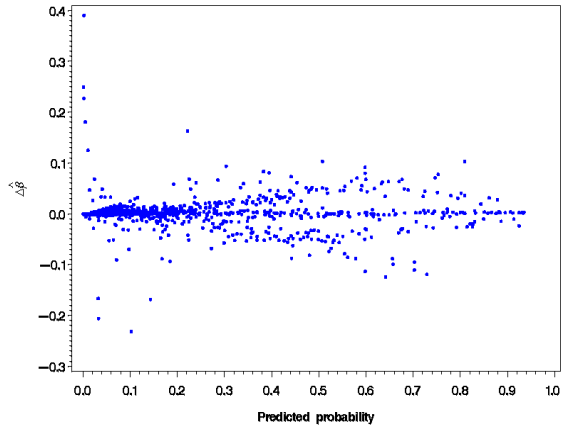
Figure H.2 DFBETAs for Reduced Model



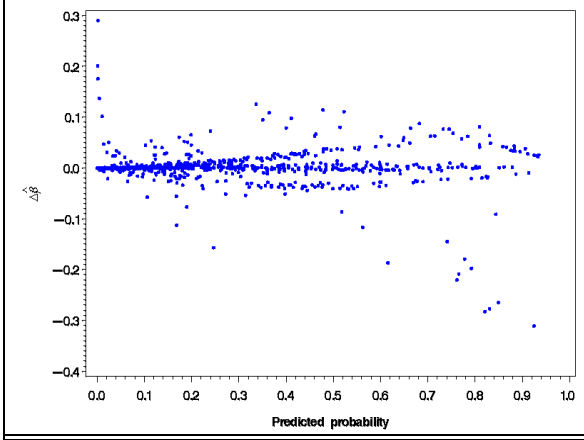
Base excess -7.0 to –9.9



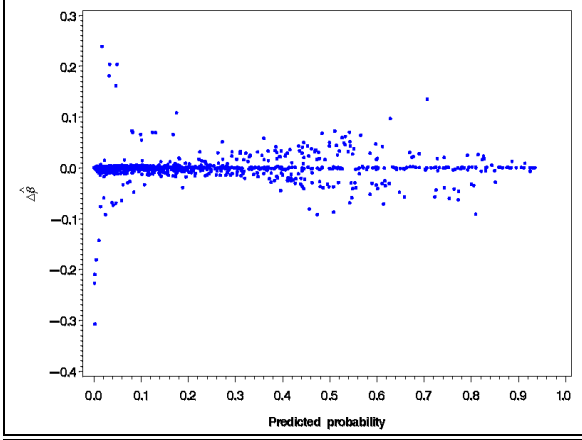
Base excess -10.0 to –14.9



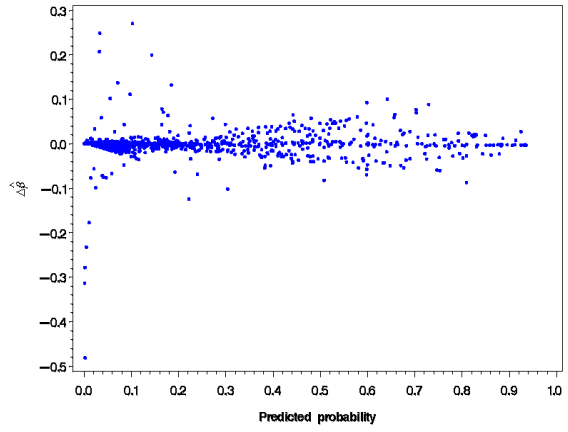
Base excess <-15



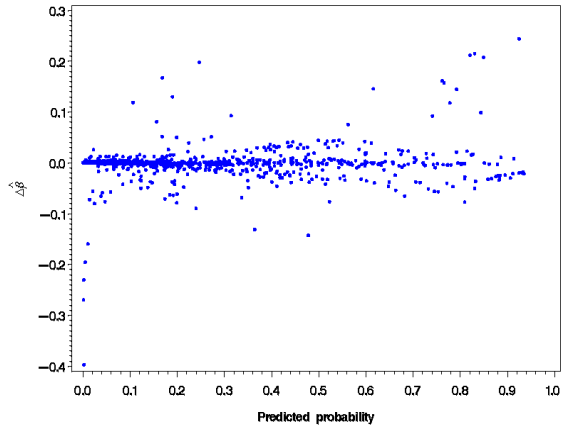
Base excess -7.0 to –9.9*weight



Base excess -10.0 to –14.9* weight



Base excess <-15* weight



Appendix I: BAYESIAN RISK-ADJUSTED MODEL

This Appendix shows WinBUGS code and diagnostic plots for the Bayesian modelling of the 'reduced' model (§6.10). First, code and plots are shown for the model to estimate the parameter values using all of the data: the results of this analysis are shown in Table 6.9. Second, an example of the code used to estimate the SMRs for each unit is reproduced, using Unit 1 as the example.

Appendix I.1 Estimation of risk-adjustment parameter values using all of the data

This code, and the subsequent plots, estimate the parameter estimates for the 'reduced' model described in §6.10.

```
model reduced {

  for (i in 1:2885) {

  died[i] ~ dbern(p[i])

c_gest[i] <- gest[i]-30
kg_bwt[i] <- (bwt[i]/1000)-1.5

logit(p[i]) <-    beta.int
                  + beta.g*c_gest[i]
                  + beta.s*gender[i]
                  + beta.a*apgar1[i]
                  + beta.w*kg_bwt[i]
                  + beta.ww*kg_bwt[i]*kg_bwt[i]
                  + beta.bg2*bg2[i]
                  + beta.bg3*bg3[i]
                  + beta.bg4*bg4[i]
                  + beta.bg2.w*bg2[i]*kg_bwt[i]
                  + beta.bg3.w*bg3[i]*kg_bwt[i]
                  + beta.bg4.w*bg4[i]*kg_bwt[i]
 }

beta.int ~ dnorm(0,1.0E-6)
beta.g ~ dnorm(0,1.0E-6)
beta.s ~ dnorm(0,1.0E-6)
beta.a ~ dnorm(0,1.0E-6)
beta.w ~ dnorm(0,1.0E-6)
beta.ww ~ dnorm(0,1.0E-6)
beta.bg2 ~ dnorm(0,1.0E-6)
beta.bg3 ~ dnorm(0,1.0E-6)
beta.bg4 ~ dnorm(0,1.0E-6)
beta.bg2.w ~ dnorm(0,1.0E-6)
```

```
beta.bg3.w ~ dnorm(0,1.0E-6)
beta.bg4.w ~ dnorm(0,1.0E-6)

}
```

Figure I.1 Brooks-Gelman-Ruben statistic plots

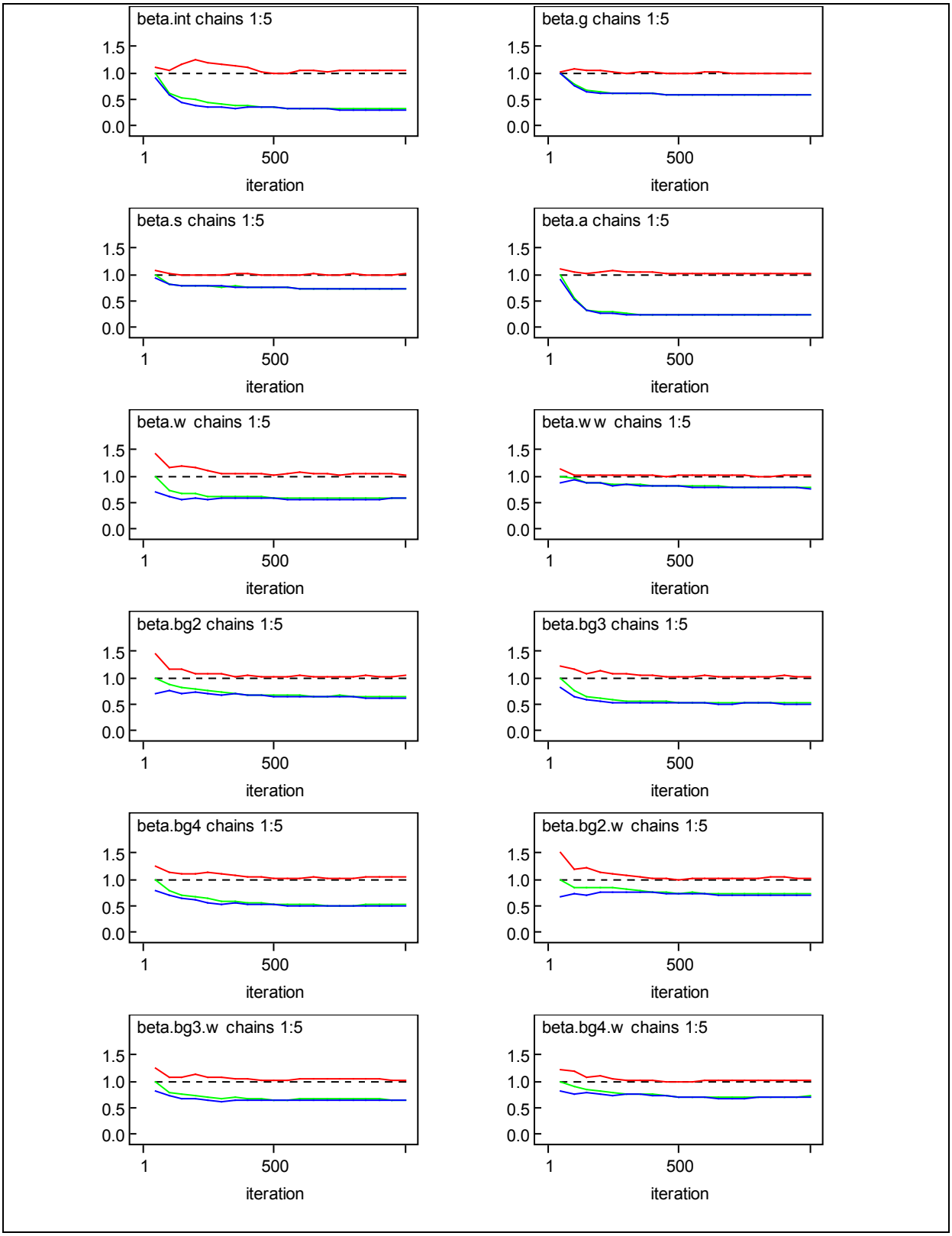
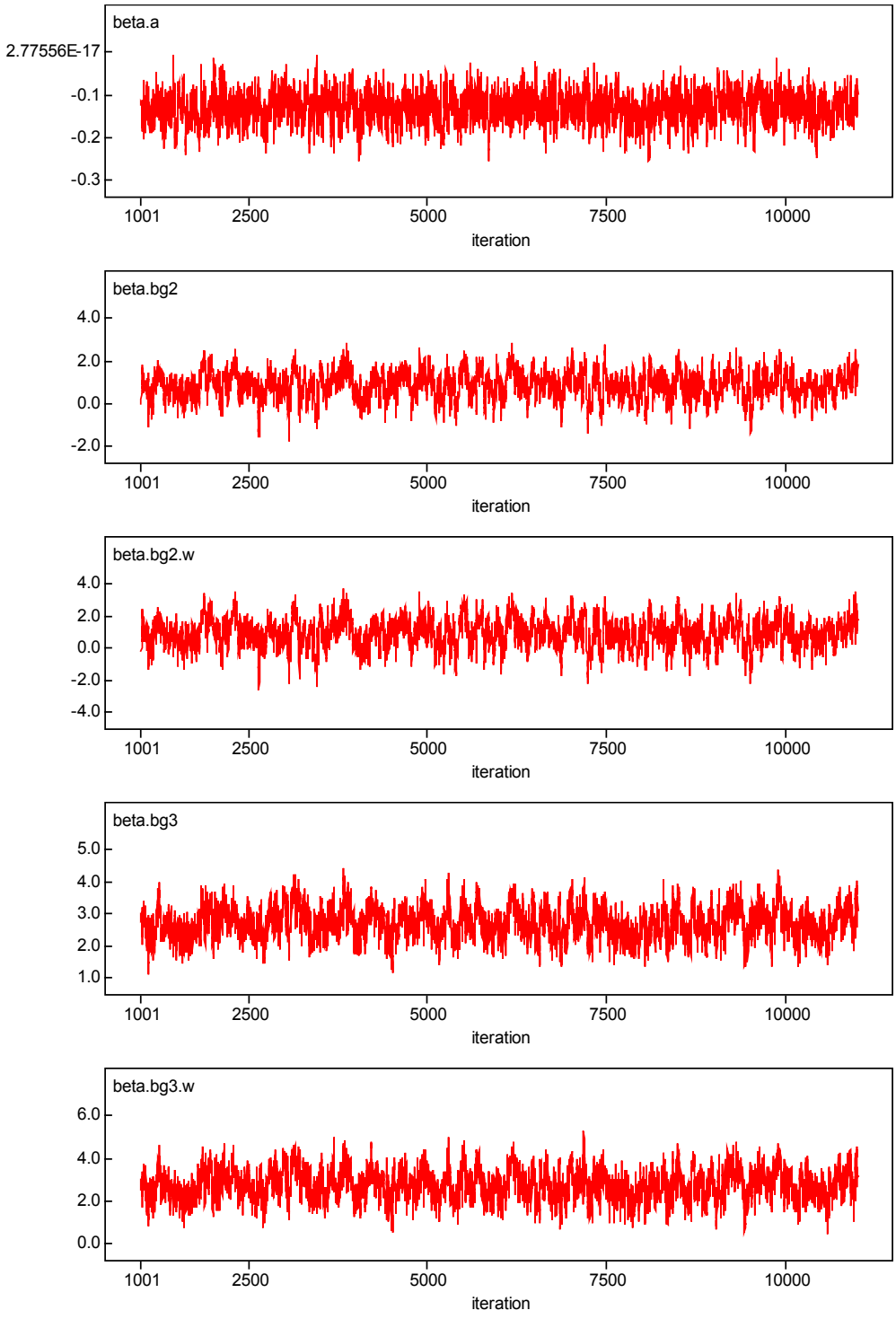
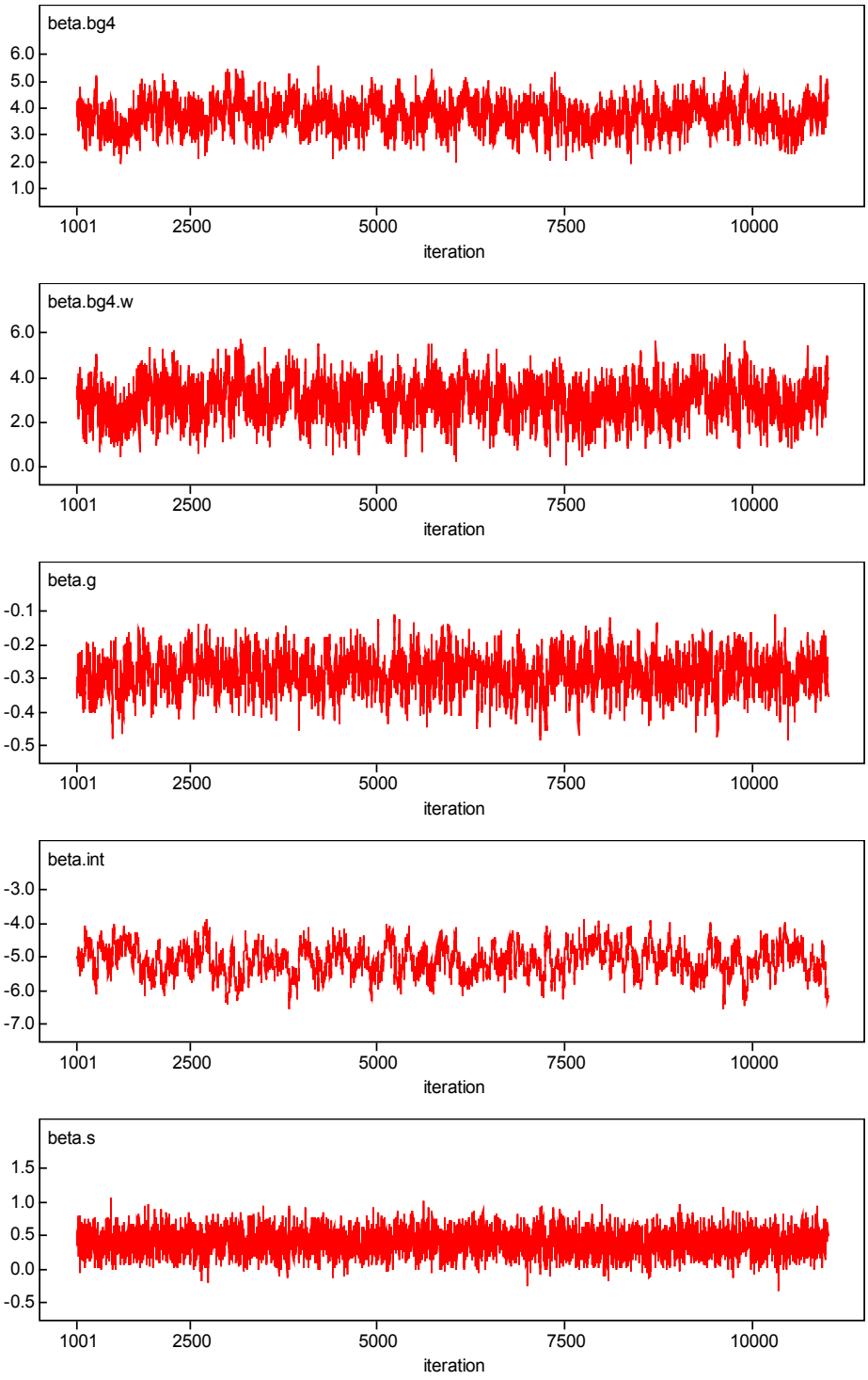


Figure I.2 Trace plots




beta.bg4
6.0
5.0
4.0
3.0
2.0
1.0
1001
2500
5000
7500
10000
iteration
beta.bg4.w
6.0
4.0
2.0
0.0
1001
2500
5000
7500
10000
iteration
beta.g
-0.1
-0.2
-0.3
-0.4
-0.5
1001
2500
5000
7500
10000
iteration
beta.int
-3.0
-4.0
-5.0
-6.0
-7.0
1001
2500
5000
7500
10000
iteration
beta.s
1.5
1.0
0.5
0.0
-0.5
1001
2500
5000
7500
10000
iteration

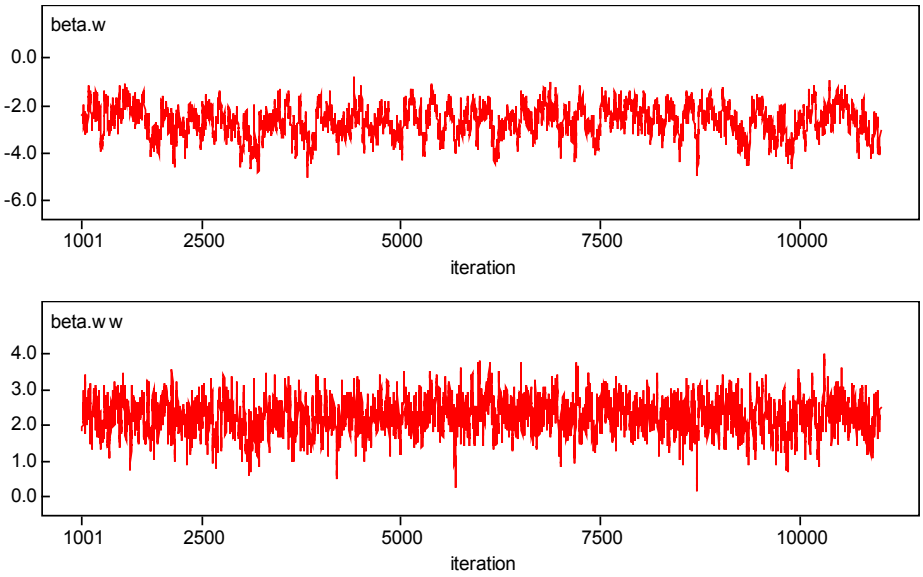

beta.w
0.0
-2.0
-4.0
-6.0
1001
2500
5000
7500
10000
iteration
beta.w w
4.0
3.0
2.0
1.0
0.0
1001
2500
5000
7500
10000
iteration

Figure I.3 Density plots

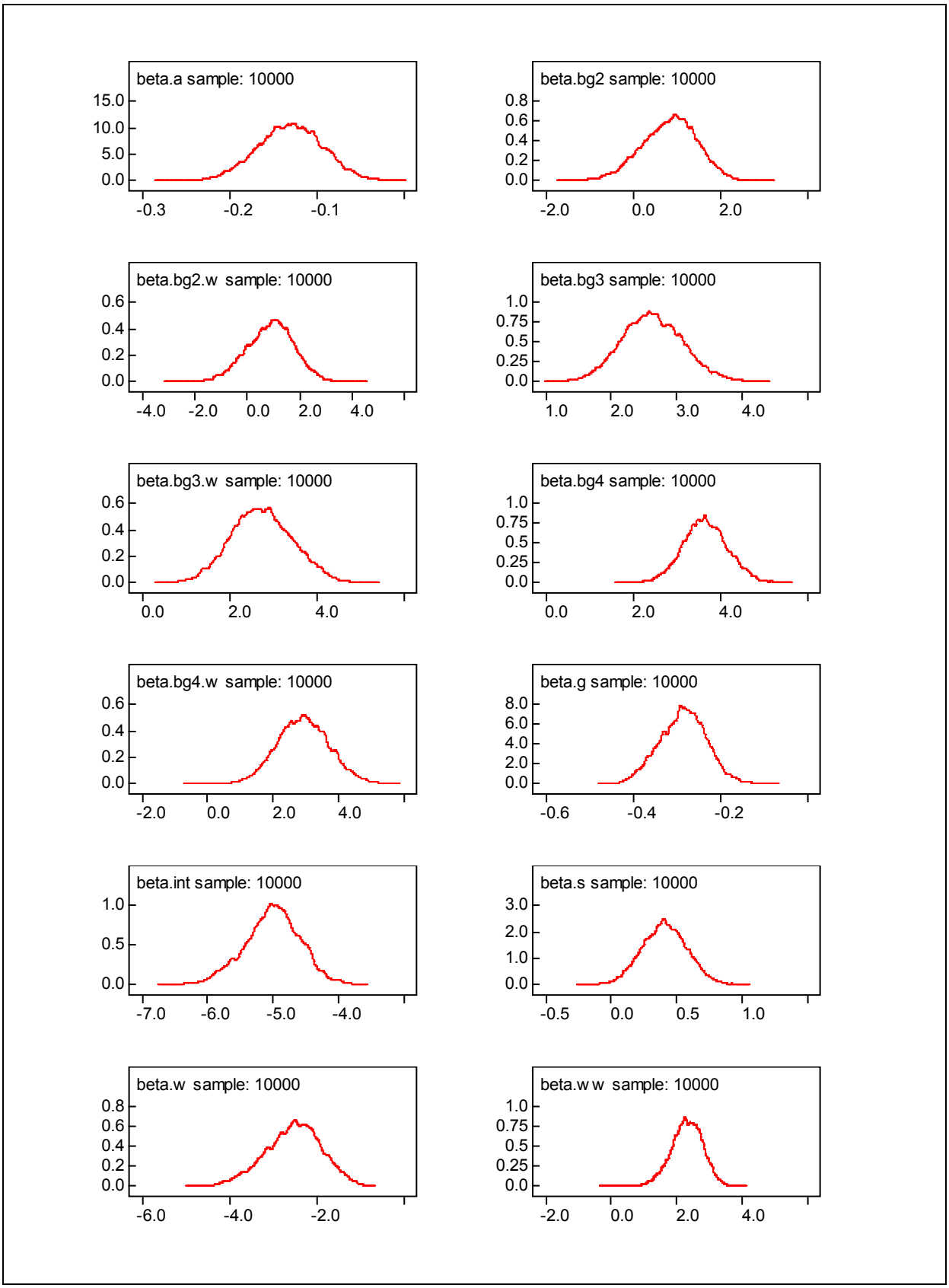
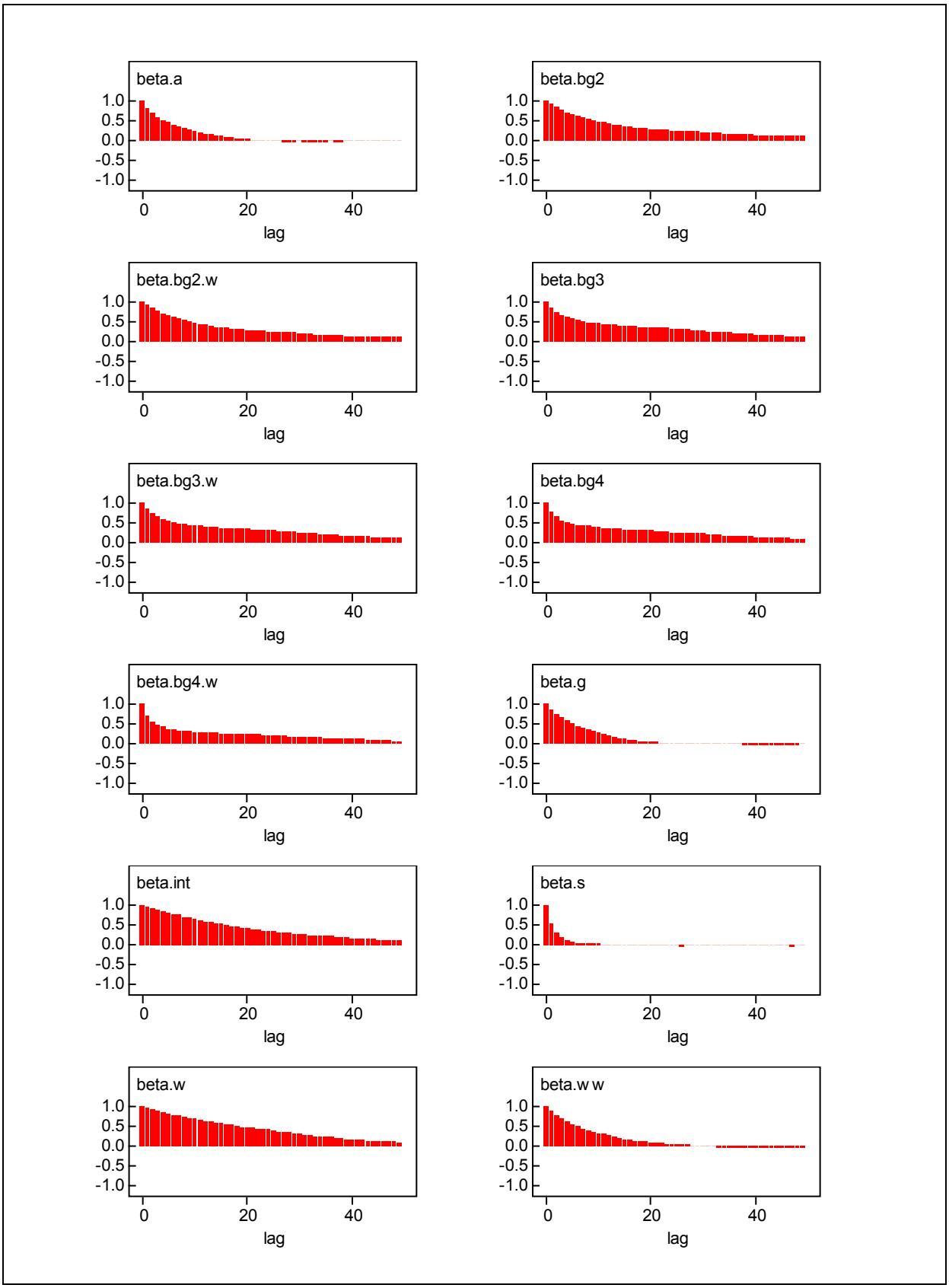


Figure I.4 Auto-correlation plots



Appendix I.2 Estimation of unit specific SMR

The WinBUGS code below is an example of the code used to estimate unit specific SMRs in §6.10. Here, the code for Unit 1 is shown.

```
model ci {for (i in 206:2853) {     # Reference data

died[i] ~ dbern(p[i])

c_gest[i] <- gest[i]-30
kg_bwt[i] <- (bwt[i]/1000)-1.5

# Estimate model parameters from reference data
logit(p[i]) <-    beta.int + risk[i]
                  + b2*i2[i]+ b3*i3[i]+ b4*i4[i]+ b5*i5[i]
                  + b6*i6[i]+ b7*i7[i]+ b8*i8[i]+ 10*i10[i]
                  + b11*i11[i]+ b12*i12[i]+ b13*i13[i]
                  + b14*i14[i]+ b15*i15[i]

                  risk[i] <-  beta.g*c_gest[i]
                              + beta.s*gender[i]
                              + beta.a*apgar1[i]
                              + beta.w*kg_bwt[i]
                              + beta.ww*kg_bwt[i]*kg_bwt[i]
                              + beta.bg2*bg2[i]
                              + beta.bg3*bg3[i]
                              + beta.bg4*bg4[i]
                              + beta.bg2.w*bg2[i]*kg_bwt[i]
                              + beta.bg3.w*bg3[i]*kg_bwt[i]
                              + beta.bg4.w*bg4[i]*kg_bwt[i]

}


for (i in 1:205)  {           # Data from unit of interest

c_gest[i] <- gest[i]-30
kg_bwt[i] <- (bwt[i]/1000)-1.5

# Calculate expected 'p' using parameters estimated above

logit(pp[i]) <-   beta.int
                  + beta.g*c_gest[i]
                  + beta.s*gender[i]
                  + beta.a*apgar1[i]
                  + beta.w*kg_bwt[i]
                  + beta.ww*kg_bwt[i]*kg_bwt[i]
                  + beta.bg2*bg2[i]
                  + beta.bg3*bg3[i]
                  + beta.bg4*bg4[i]
                  + beta.bg2.w*bg2[i]*kg_bwt[i]
                  + beta.bg3.w*bg3[i]*kg_bwt[i]
                  + beta.bg4.w*bg4[i]*kg_bwt[i]


# Estimate model parameters from unit of interest

died[i] ~ dbern(op[i])
```

```
op[i] <- exp(logit.op[i])/(1+exp(logit.op[i]))
logit.op[i] <-   ibeta.int
                  + ibeta.g*c_gest[i]
                  + ibeta.s*gender[i]
                  + ibeta.a*apgar1[i]
                  + ibeta.w*kg_bwt[i]
                  + ibeta.ww*kg_bwt[i]*kg_bwt[i]
                  + ibeta.bg2*bg2[i]
                  + beta.bg3*bg3[i]
                  + beta.bg4*bg4[i]
                  + ibeta.bg2.w*bg2[i]*kg_bwt[i]
                  + ibeta.bg3.w*bg3[i]*kg_bwt[i]
                  + ibeta.bg4.w*bg4[i]*kg_bwt[i]

}

# Calculate SMR
sum.pp <- sum(pp[])                 # Sum of predicted

sum.ob <- sum(op[])                 # Sum of observed

ratio <- s.ob/sum.pp                # SMR

over <- step(ratio-1.5)
under <- step(0.666667-ratio)

# Prior distributions
beta.int ~ dnorm(-4.6,1.0E-6)
beta.g ~ dnorm(0,1.0E-6)
beta.s ~ dnorm(0,1.0E-6)
beta.a ~ dnorm(0,1.0E-6)
beta.w ~ dnorm(0,1.0E-6)
beta.ww ~ dnorm(0,1.0E-6)
beta.bg2 ~ dnorm(0,1.0E-6)
beta.bg3 ~ dnorm(0,1.0E-6)
beta.bg4 ~ dnorm(0,1.0E-6)
beta.bg2.w ~ dnorm(0,1.0E-6)
beta.bg3.w ~ dnorm(0,1.0E-6)
beta.bg4.w ~ dnorm(0,1.0E-6)

b1 ~ dnorm(0,1)
b2 ~ dnorm(0,1)
b3 ~ dnorm(0,1)
b4 ~ dnorm(0,1)
b5 ~ dnorm(0,1)
b6 ~ dnorm(0,1)
b7 ~ dnorm(0,1)
b8 ~ dnorm(0,1)
b9 ~ dnorm(0,1)
b10 ~ dnorm(0,1)
b11 ~ dnorm(0,1)
b12 ~ dnorm(0,1)
b13 ~ dnorm(0,1)
b14 ~ dnorm(0,1)
b15 ~ dnorm(0,1)

beta.int.c <- cut(beta.int)
beta.g.c <- cut(beta.g)
beta.s.c <- cut(beta.s)
beta.a.c <- cut(beta.a)
beta.w.c <- cut(beta.w)
```

```
        beta.ww.c <- cut(beta.ww)
        beta.bg2.c <- cut(beta.bg2)
        beta.bg3.c <- cut(beta.bg3)
        beta.bg4.c <- cut(beta.bg4)
        beta.bg2.w.c <- cut(beta.bg2.w)
        beta.bg3.w.c <- cut(beta.bg3.w)
        beta.bg4.w.c <- cut(beta.bg4.w)

        ibeta.int ~ dnorm(beta.int.c,1.0E-1)
        ibeta.g ~ dnorm(beta.g.c,1.0E-1)
        ibeta.s ~ dnorm(beta.s.c,1.0E-1)
        ibeta.a ~ dnorm(beta.a.c,1.0E-1)
        ibeta.w ~ dnorm(beta.w.c,1.0E-1)
        ibeta.ww ~ dnorm(beta.ww.c,1.0E-1)
        ibeta.bg2 ~ dnorm(beta.bg2.c,1.0E-1)
        ibeta.bg3 ~ dnorm(beta.bg3.c,1.0E-1)
        ibeta.bg4 ~ dnorm(beta.bg4.c,1.0E-1)
        ibeta.bg2.w ~ dnorm(beta.bg2.w.c,1.0E-1)
        ibeta.bg3.w ~ dnorm(beta.bg3.w.c,1.0E-1)
        ibeta.bg4.w ~ dnorm(beta.bg4.w.c,1.0E-1)
}
```

Appendix J: ADDITIONAL GRAPHICS

In this Appendix two plots are reproduced that have been used to compare hospitals. The first is an example of **Radar plots** proposed by Leary et al. (2002) and the second is Florence Nightingale's **Coxcomb**.

Figure J.1 Radar plots

(Leary et al. 2002)

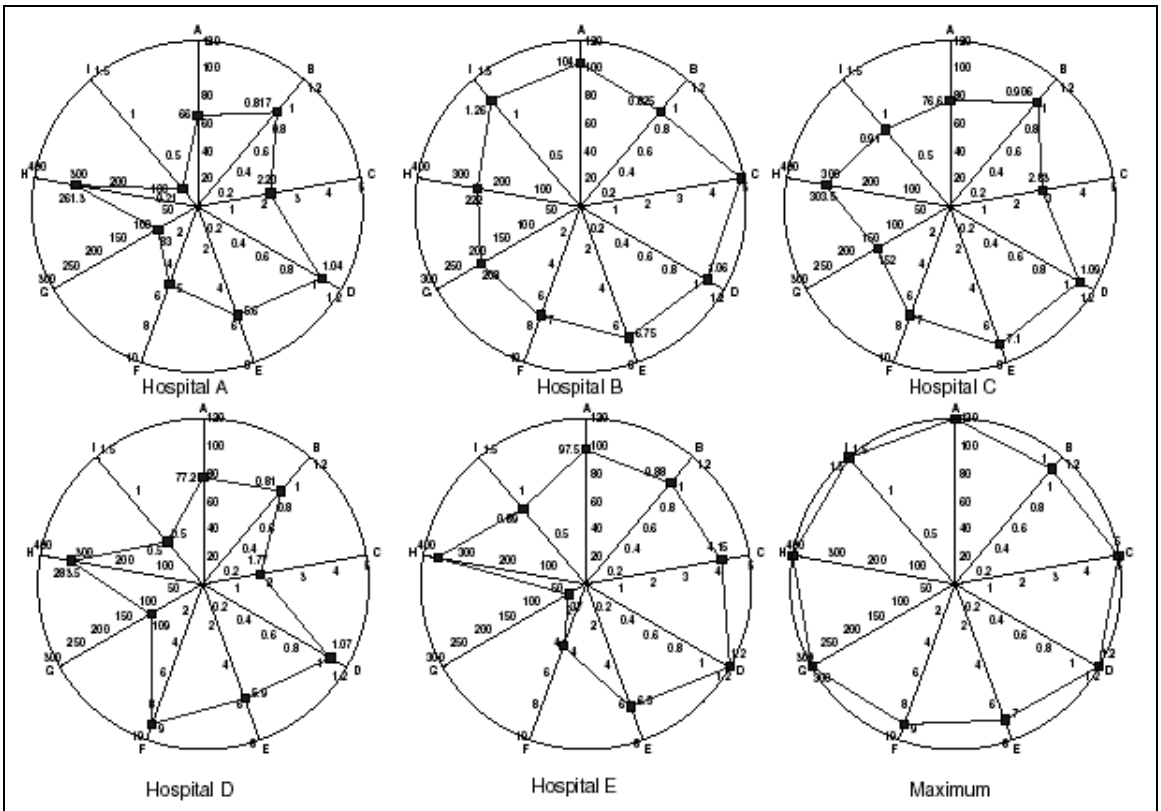
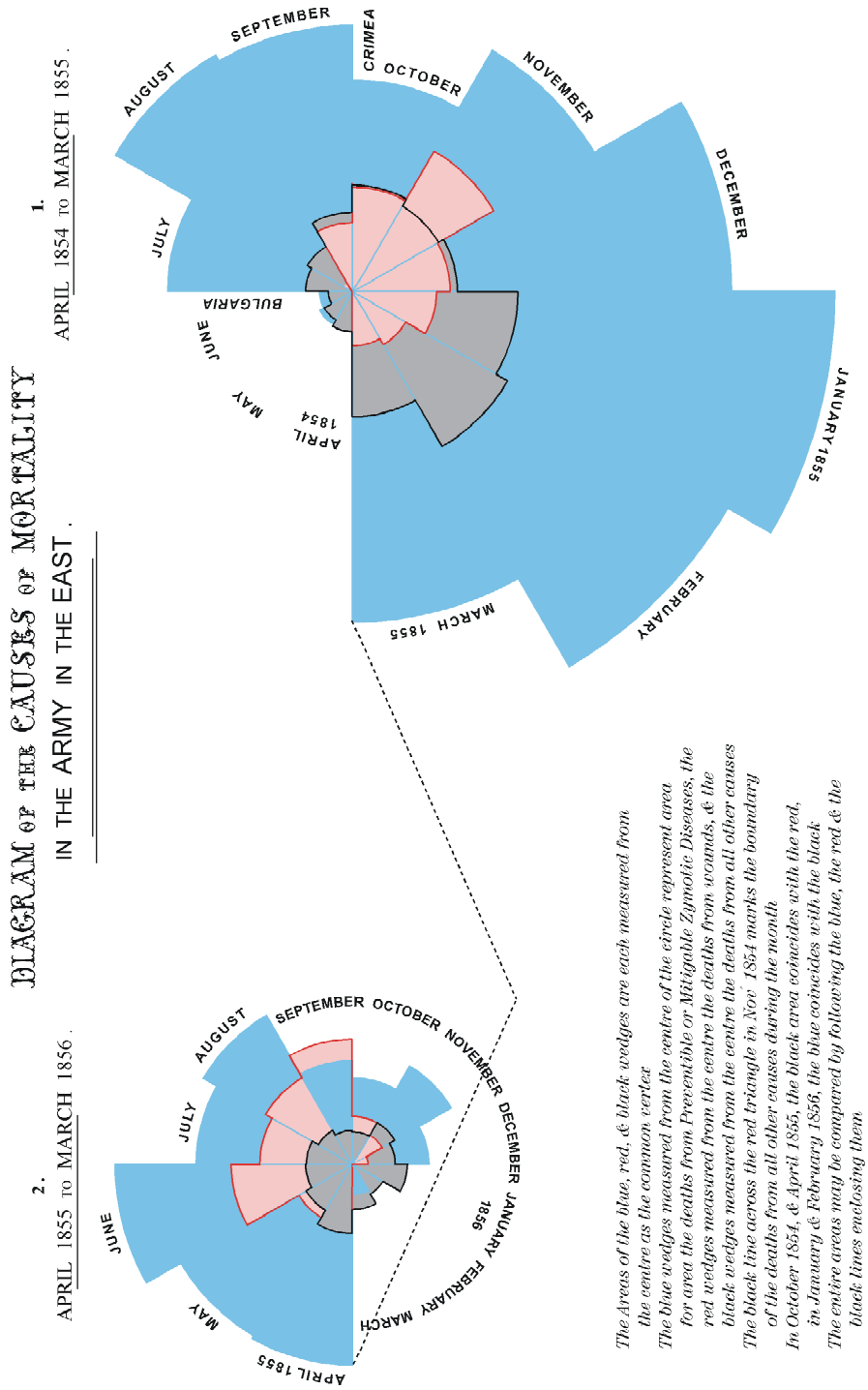


Figure 1 The radar charts for the five critical care units and the maximum available. A: Total number patients admitted per critical care bed per year. B: Occupancy. C: Actual percentage of patients transferred for non-clinical reasons subtracted from the upper limit of 5%. D: Transformed SMR = 1 + (1 – actual SMR). E: Whole time equivalents of nursing staff per critical care bed. F: Number of protocols available. G: The national average cost per patient-day minus the average cost per patient-day for ICU/HDU. H: Total number patient-days of care provided per critical care bed per year. I: The national average length of stay minus the average length of stay in the ICU/HDU.

Figure J.2 Coxcomb

(Reproduced from www.florence-nightingale-avenging-angel.co.uk/Coxcomb.htm.)



REFERENCES

Adab, P., Rouse, A.M., Mohammed, M.A. and Marshall, T. (2002) Performance league tables: the NHS deserves better. *British Medical Journal* **324**:95-98.

AFD Software Ltd. (2004)
Available at http://www.afd.co.uk/pcplus.asp

Agresti, A. (1990) Categorical Data Analysis. New York, Wiley.

Agresti, A. (1993) Distribution-free fitting of logit models with randon effects for repeated categorical responses. *Statistics in Medicine* **12**:1969-1987.

Agresti, A., Caffo, B. and Ohman-Strickland, P. (2004) Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis* **47**:639-653.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalised linear models. *Statistical Computing* **6**:251-262.

Al-Azzawi, F. (1990) A colour atlas of childbirth and obstetric techniques. London, Wolfe.

Alberman, E. (1991) Are our babies becoming bigger? *Journal of the Royal Society of Medicine* **84**:257-260.

Albert, A. and Anderson, J.A. (1984) On the existence of maximum-likelihood estimates in logistic- regression models. *Biometrika* **71**:1-10.

Alexander, G.R., Himes, J.H., Kaufman, R.B., Mor, J. and Kogan, M. (1996) A United States national reference for fetal growth. *Obstetrics and Gynecology* **87**:163-168.

Allen, M.C. (2002) Preterm outcomes research: a critical component of neonatal intensive care. *Mental Retardation and Developmental Disabilities Research Reviews* **8**:221-233.

Altman, D.G. and Chitty, L.S. (1997) New charts for ultrasound dating in pregnancy. *Ultrasound in Obstetrics and Gynecology* **10**:174-191.

Altman, D.G. and Royston, P. (2000) What do we mean by validating a prognostic model? *Statistics in Medicine* **19**:453-473.

Amaru, R.C., Bush, M.C., Berkowitz, R.L., Lapinski, R.H. and Gaddipati, S. (2004) Is discordant growth in twins an independent risk factor for adverse neonatal outcome? *Obstetrics and Gynecology* **103**:71-76.

Ambler, G., Omar, R.Z. and Royston, P. (2005) A comparison of methods for handling missing predictor values in prognostic models. 26th Annual Conference The International Society for Clinical Biostatisticians.

Apgar, V. (1953) Proposal for a new method of evaluation of newborn infants. *Current Researches in Anesthesia and Analgesia* **32**:260-267.

Arbuckle, T.E., Wilkins, R. and Sherman, G.J. (1993) Birth-weight percentiles by gestational-age in Canada. *Obstetrics and Gynecology* **81**:39-48.

Armitage, P. and Berry, G. (1994) Statistical Methods in Medical Research. 3rd edn, Oxford, Blackwell Science.

Arnold, C. (1992) Very low birth weight: a problematic cohort for epidemiologic studies of very small or immature neonates. *American Journal of Epidemiology* **136**:767-768.

Arnold, C.C., Kramer, M.S., Hobbs, C.A., McLean, F.H. and Usher, R.H. (1991) Very low birth weight: a problematic cohort for epidemiologic studies of very small or immature neonates. *American Journal of Epidemiology* **134**:604-613.

Audit Commission (1992) Children first: a study of hospital services. Audit Commission Services report No. 7. London, HMSO.

Austin, P.C. (2002) A Comparison of Bayesian Methods for Profiling Hospital Performance. *Medical Decision Making* **22**:163-172.

Austin, P.C., Alter, D.A. and Tu, J.V. (2003) The use of fixed- and randon-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. *Medical Decision Making* **23**:526-539.

Austin, P.C., Naylor, C.D. and Tu, J.V. (2001) A comparison of a Bayesian vs. a frequentist

method for profiling hospital performance. *Journal of Evaluation in Clinical Practice* **7**:35-45.

Aveyard, P., Manaseki, S. and Chambers, J. (2002) The relationship between mean birth weight and poverty using the Townsend deprivation score and the Super Profile classification system. *Public Health* **116**:308-314.

Aylin, P., Alves, B., Best, N., Cook, A., Elliott, P., Evans, S.J.W., Lawrence, A.E., Murray, G.D., Pollock, J. and Spiegelhalter, D. (2001a) Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-96: Was Bristol an outlier? *Lancet* **358**:181-187.

Aylin, P., Jarman, B. and Elliot, P. (2005) Paediatric cardiac surgical mortality after Bristol. *British Medical Journal* **330**:44

Aylin, P., Spiegelhalter, D., Best, N., Murray, G.D. and Elliott, P. (2001b) Was Bristol an outlier. *Lancet* **358**:2084-2084.

Ayoubi, J.-M., Audibert, F., Vial, M., Pons, J.C., Taylor, S. and Frydman, R. (2002) Fetal heart rate and survival of the very premature newborn. *American Journal of Obstetrics and Gynecology* **187**:1026-1030.

Bagust, A. (1996) League tables. *British Journal of Hospital Medicine* **55** (6):369-70.

Bai, J., Wong, F.W.S., Bauman, A. and Mohsin, M. (2002) Parity and pregnancy outcome. *American Journal of Obstetrics and Gynecology* **186**:274-278.

Baker, R., Jones, D.R. and Goldblatt, P. (2003) Monitoring mortality rates in general practice after Shipman. *British Medical Journal* **326**:274-276.

Bakketeig, L.S. and Hoffman, H.J. (1979) Perinatal mortality by birth order within cohorts based on sibship size. *British Medical Journal* **2**(6192):693-696.

Ballard, J.L., Khoury, J.C., Wedig, K., Wang, L., Eilerswalsman, B.L. and Lipp, R. (1991) New Ballard Score, expanded to include extremely premature infants. *Journal of Pediatrics* **119**:417-423.

Ballard, J.L., Novak, K.K. and Driver, M. (1979) A simplified score for assessment of fetal

maturation of newly born infants. *Journal of Pediatrics* **95**:769-774.

Bambang, S., Spencer, N.J., Logan, S. and Gill, L. (2000) Cause-specific perinatal death rates, birth weight and deprivation in the West Midlands, 1991-93. *Child Care Health and Development* **26**:73-82.

Bard, H. (1993) Assessing neonatal risk: CRIB vs SNAP. *Lancet* **342**(8869):449-450.

Bastos, G., Gomes, A., Oliveira, P. and da Silva, A.T. (1997) A comparison of 4 pregnancy assessment scales (CRIB, SNAP, SNAP-PE, NTISS) in premature newborns. Clinical Risk Index for Babies. Score for Neonatal Acute Physiology. Score for Neonatal Acute Physiology-Perinatal Extension. Neonatal Therapeutic Intervention Scoring System. *Acta Medica Portuguesa* **10** (2-3):161-165.

Bates, D.W. and Gawande, A.A. (2000) Error in medicine: what have we learned? *Annals of Internal Medicine* **132**(9):763-767.

Battin, M., Ling, E.W.Y., Whitfield, M.F., Mackinnon, M. and Effer, S.B. (1998) Has the outcome for extremely low gestational age (ELGA) infants improved following recent advances in neonatal intensive care? *American Journal of Perinatology* **15**:469-477.

Bauer, J., Hentschel, R., Zahradnik, H., Karck, U. and Linderkamp, O. (2003) Vaginal delivery and neonatal outcome in extremely-low-birth-weight infants below 26 weeks of gestational age. *American Journal of Perinatology* **20**:181-188.

Baumer, J.H., Wright, D. and Mill, T. (1997) Illness severity measured by CRIB score: a product of changes in perinatal care? *Archives of Disease in Childhood Fetal & Neonatal Edition* **77** (3):F211-F215.

Becher, E.C. and Chassin, M.R. (2001) Improving quality, minimizing error: making it happen. A five-point plan and why we need it. *Health Affairs* **20**(3):68-81.

Bednarek, F.J., Weisberger, S., Richardson, D.K., Frantz, I.D. 3rd, Shah, B. and Rubin, L.P. (1998) Variations in blood transfusions among newborn intensive care units. SNAP II Study Group. *Journal of Pediatrics* **133** (5):601-607.

Beischer, N.A. and Mackay, E.V. (1978) Obstetrics and the newborn. Eastbourne, Saunders.

Berkowitz, G.S., Skovron, M.L., Lapinski, R.H. and Berkowitz, R.L. (1990) Delayed childbearing and the outcome of pregnancy. *New England Journal of Medicine* **322**:659-664.

Berman, S., Richardson, D.K., Cohen, A.P., Pursley, D.M. and Lieberman, E. (2001) Relationship of race and severity of neonatal illness. *American Journal of Obstetrics and Gynecology* **184**:668-672.

Berry, D.A. and Stangl, D.K. (1996) Bayesian methods in health-related research. In: Berry, D.A. and Stangl, D.K., (Eds.), *Bayesian Biostatistics*. New York, Marcel Dekker: pp. 3-66.

Berwick, D.M. (2001) Not again! Preventing errors lies in redesign – not exhortation. *British Medical Journal* **322**:247-248.

Best, N., Cowles, M.K. and Vines, K. (1996) CODA Convergence diagnosis and output software for Gibbs sampling output: version 0.30. Cambridge, MRC Biostatistics Unit.

Bhopal, R. (2002) Concepts of epidemiology: an integrated introduction to the ideas, theories, principles, and methods of epidemiology. Oxford, Oxford University Press.

Billewicz, W.Z. (1973) Some implications of self-selection for pregnancy. *British Journal of Preventative and Social Medicine* **27**:49-52.

Bland, M. (1995) An introduction to medical statistics. 2nd edn, Oxford, Oxford University Press.

Blondel, A., Morin, I., Platt, R.W., Kramer, M.S., Usher, R. and Breart, G. (2002) Algorithms for Combining Menstrual and Ultrasound Estimates of Gestational Age: Consequences for Rates of Preterm and Postterm Birth. *BJOG - an International Journal of Obstetrics and Gynaecology* **109**:718-720.

Bloomfield, F.H., Oliver, M.H., Hawkins, P., Campbell, M., Phillips, D.J., Gluckman, P.D., Challis, J.R.G. and Harding, J.E. (2003) A periconceptional nutritional origin for noninfectious preterm birth. *Science* **300**:606

Bohin, S., Draper, E.S. and Field, D.J. (1999) Health status of a population of infants born before 26 weeks gestation derived from routine data collected between 21 and 27 months post-delivery. *Early Human Development* **55**:9-18.

Bosman, R.J., Oudemans van Straaten, H.M. and Zandstra, D.F. (1998) The use of intensive care information systems alters outcome prediction. *Intensive Care Medicine* **24**:953-958.

Boyd, C.R., Tolson, M.A. and Copes, W.S. (1987) Evaluating trauma care: The TRISS method. *Journal of Trauma* **27**:370

Boyd, O. and Grounds, M. (1994) Can standardized mortality ratio be used to compare quality of intensive-care unit performance. *Critical Care Medicine* **22**:1706-1708.

Bradley, R.A. and Terry, M.A. (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**:324-335.

Braitman, L.E. and Davidoff, F. (1996) Predicting clinical states in individual patients. *Annals of Internal Medicine* **125** (5):406-412.

Breslow, N.E. and Day, N.E. (1987) Statistical methods in cancer research: Volume II - the design and analysis of cohort studies. Oxford, International Agency for Research in Cancer.

Bridgewater, B., Hooper, T., Campbell, C., Jones, M., Carey, J., Waterworth, P., Deiraniya, A. and Yonan, N. (2002) Performance league tables: publication of league tables needs to be open and accurate. *British Medical Journal* **324**:542-543.

British Association of Perinatal Medicine (2001) Standards for hospitals providing neonatal intensive and high dependency care. London, BAPM.

Brooks, S.P. & Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455

Bryson, M.C. and Johnson, M.E. (1981) The incidence of monotone likelihood in the Cox model. *Technometrics* **23**:381-384.

Burton, P.R., Gurrin, L.C. and Campbell, M.J. (1998) Clinical significance not statistical significance: a simple Bayesian alternative to p values. *Journal of Epidemiology & Community Health* **52**(5):318-323.

Butler, S.M. and Louis, T.A. (1992) Random effects models with non-parametric priors. *Statistics in Medicine* **11**:1981-2000.

Campbell, S. and Newman, G.B. (1971) Growth of the fetal biparietal diameter during normal pregnancy. *Journal of Obstetrics and Gynaecology of the British Commonwealth* **78**:513-519.

Carr, R. (2000) Neutrphil production and function in newborn infants. *British Journal of Haematology* **110**:18-28.

Casey, B.M., Mcintire, D.D. and Leveno, K.J. (2001) The continuing value of the Apgar score for the assessment of newborn infants. *New England Journal of Medicine* **344**:467-471.

Center for Health Services Research in Primary Care (1996) Second Report of the California Hospital Outcomes Project (1996): Acute Myocardial Infarction Volume Two: Technical Appendix: Chapter 008. Sacramento, CA, California Office of Statewide Health Planning and Development.

CESDI (2003) Project 27/28. An enquiry into quality of care and its effect on the survival of babies born at 27-28 weeks: Executive summary. Norwich, TSO.

Chaiken, B.P. (1996) Impact on provider profiling. *Virginia Medical Quarterly* **123**(4):238-240.

Chaix, B., Bobashev, G., Merlo, J. and Chauvin, P. (2004) Re: Detecting patterns of occupational illness clustering with alternating logistic regressions applied to longitudinal data. *American Journal of Epidemiology* **160**(5):505-506.

Chappell, K. and Newman, C. (2003) Potential tenfold drug overdoses on a neonatal unit. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89:**F483-F484

Chernick, M.R. (1999) Bootstrap methods: a practitioner's guide. New York, Wiley.

Cheung, Y.B., Yip, P. and Karlberg, J. (2000) Mortality of twins and singletons by gestational age: a varying-coefficient approach. *American Journal of Epidemiology* **152**:1107-1116.

Chiswick, M. (1995) Antenatal TRH. *Lancet* **345**:877-882.

Christiansen, C.L. and Morris, C.N. (1997) Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* **127**(8 Pt 2):764-768.

Clark, D.A. and Hakanson, D.O. (1988) The inaccuracy of Apgar scoring. *Journal of Perinatology* **8**(3):203-205.

Clarke, P.A. (2001) What residents are not learning: observations in an NICU. *Academic Medicine* **76**(5):419-424.

Classen, D.C. and Kilbridge, P.M. (2002) The roles and responsibility of physicians to improve patient safety within health care delivery systems. *Academic Medicine* **77**(10): 963-972.

Cohen, J. (1994) The Earth is round (p less than .05). *American Psychologist* **49**:997-1003.

Cohen, S.B., Dulitzky, M., Lipitz, S., Mashiach, S. and Schiff, E. (1997) New birth weight nomograms for twin gestation on the basis of accurate gestational age. *American Journal of Obstetrics and Gynecology* **177**:1101-1104.

Cole, T.J. (1993) Scaling and rounding regression coefficients to integers. *Applied Statistics* **42**:461-468.

Cole, T.J., Morley, C.J., Thornton, A.J. and Fowler, M.A. (1991) A scoring system to quantify illness in babies under 6 months of age. *Journal Of The Royal Statistical Society Series A - Statistics In Society* **154**:287-304.

Colver, A.F., Gibson, M., Hey, E.N., Jarvis, S.N., Mackie, P.C. and Richmond, S. (2000) Increasing rates of cerebral palsy across the severity spectrum in north-east England 1964-1993. *Archives of Disease in Childhood Fetal & Neonatal Edition* **83**:F7-F12

Committee on Information Technology in Medicine (1991) Hospital-specific mortality rates for cardiac surgery. *New York State Journal of Medicine* **91**(10):461-462.

Cooper, R.L., Goldenberg, R.L., Creasy, R.K., DuBard, M.B., Davis, R.O., Entman, S.S., Iams, J.D. and Cliver, S.P. (1993) A multicenter study of preterm birth weight and gestational age-specific neonatal mortality. *American Journal of Obstetrics and Gynecology* **168**:78-84.

Coory, M. (1997) Does gestational age in combination with birthweight provide better statistical adjustment of neonatal mortality rates than birthweight alone? *Paediatric & Perinatal Epidemiology* **11**(4):385-391.

Court, B.V. and Cheng, K.K. (1995) Pros and cons of standardised mortality ratios. *Lancet* **346**:1432.

Cox, D.R. (1958) Two further applications of a model for binary regression. *Biometrika* **45**:562-565.

Creasy, R.K. (1997) Management of labor and delivery. Malden Massachusetts, Blackwell Science.

Crouchley, R. and Taylor, J. (2004) Higher education performance indicators: invited comments on the papers by Draper and Gittoes and Bratti *et al*. *Journal Of The Royal Statistical Society Series A - Statistics In Society* **167**:497-498.

Crowley, P. (2003) Prophylactic corticosteroids for preterm birth (Cochrane Review). In: *The Cochrane Library, Issue 3*, Oxford, Update Software

Cullen, D.J., Civetta, J.M., Briggs, B.A. and Ferrara, L.C. (1974) Therapeutic intervention scoring system: a method for quantitative comparison of patient care. *Critical Care Medicine* **2**:57-60.

Cuttini, M., Casotto, V., Kaminski, M., de Beaufort, I., Berbik, I., Hansen, G., Kollée, L., Kucinskas, A., Lenoir, S., Levin, A., Orzalesi, M., Persson, J., Rebagliato, M., Reid, M., Saracci, R. and and other members of the EURONIC Study Group (2004) Should euthanasia be legal? An international survey of neonatal intensive care units staff. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F19-F24

Daley, J., Iezzoni, L.I. and Khuri, S.F. (1995) Complication rates as a measure of quality of care. *Journal of the American Medical Association* **274**(21):1674-1675.

Daley, J., Jencks, S., Draper, D., Lenhart, G., Thomas, N. and Walker, J. (1988) Predicting hospital-associated mortality for Medicare patients. A method for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. *Journal of the American Medical Association* **260**(24):3617-3624.

Daley, J., Khuri, S.F., Henderson, W., Hur, K., Gibbs, J.O., Barbour, G., Demakis, J., Irvin, G. 3rd, Stremple, J.F., Grover, F., McDonald, G., Passaro, E. Jr, Fabri, P.J., Spencer, J., Hammermeister, K., Aust, J.B. and Oprian, C. (1997) Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study. *Journal of the American College of Surgeons* **185** (4):328-340.

David, R.J. (1983) Population-based intrauterine growth-curves from computerized birth Certificates. *Southern Medical Journal* **76**:1401-1406.

Davidian, M. and Gallant, A.R. (1993) The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**:475-488.

Davies, H.T.O., Crombie, I.K. and Tavakoli, M. (1998) When can odds ratios mislead? *British Medical Journal* **316**:989-991.

Davison, A.C. and Hinkley, D.V. (1997) Bootstrap methods and their application. Cambridge, Cambridge University Press.

Dawes, R.M. (1980) You can't systematize human judgment: dyslexia. *New Directions for Methodology of Social and Behavioral Science* **4**:67-78.

Dawes, R.M. and Corrigan, B. (1974) Linear models in decision making. *Psychological Bulletin* **81**:95-106.

de Courcy-Wheeler, R.H., Wolfe, C.D., Fitzgerald, A., Spencer, M., Goodman, J.D. and Gamsu, H.R. (1995) Use of the CRIB (Clinical Risk Index for Babies) score in prediction of neonatal mortality and morbidity. *Archives of Disease in Childhood Fetal & Neonatal Edition* **73** (1):F32-F36.

Delamothe, T. (1998) Who killed Cock Robin? *British Medical Journal* **316**:1757

DeLong, E.R., Peterson, E.D., DeLong, D.M., Muhlbaier, L.H., Hackett, S. and Mark, D.B. (1997) Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* **16** (23):2645-2664.

Department of Health (2000) Learning from Bristol. The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995. London, The Stationery Office.

Department of Health (2002a) Learning from Bristol: The Department of Health's response to the report of the public enquiry into children's heath surgery at the Bristol Royal Infirmary 1984-1995. London, HMSO.

Department of Health (2002b) Reforming NHS financial flows: introducing payment by results. London, DOH.

Department of the Environment Transport and the Regions (2000) Indices of Deprivation

2000. London, Department of the Environment, Transport and the Regions.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (1998) Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal Of The Royal Statistical Society Series C - Applied Statistics* **47**(Pt4):511-525.

Dole, N., Savitz, D.A., Hertz-Picciotto, I., Siega-Riz, A.M., McMahon, M.J. and Buekens, P. (2003) Maternal stress and preterm birth. *American Journal of Epidemiology* **157**:14-24.

Donabedian, A. (1966) Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly* **44**:166-206.

Donabedian, A. (1978) The quality of medical care. *Science* **200**:856-864.

Donabedian, A. (1988) The quality of care: how can it be assessed. *Journal of the American Medical Association* **260**:1743-1748.

Donovan, E.F., Tyson, J.E., Ehrenkranz, R.A., Verter, J., Wright, L.L., Korones, S.B., Bauer, C.R., Shankaran, S., Stoll, B.J., Fanaroff, A.A., Oh, W., Lemons, J.A., Stevenson, D.K. and Papile, L.A. (1999) Inaccuracy of ballard scores before 28 weeks' gestation. *Journal of Pediatrics* **135**:147-152.

Dorling, J.S., Field, D.J. and Manktelow, B. (2005) Illness severity scoring systems. *Archives of Disease in Childhood Fetal & Neonatal Edition* **90**:F11-F16

Dr Foster (2004)
Available at http://www.drfoster.co.uk/.

Draper, D. and Gittoes, M. (2004) Statistical analysis of performance indicators in UK higher education. *Journal Of The Royal Statistical Society Series A - Statistics In Society* **167**:449-474.

Draper, E.S., Manktelow, B., Field, D.J. and James, D. (1999) Prediction of survival for preterm births by weight and gestational age: restrospective population based study. *British Medical Journal* **319**:1093-1097.

Draper, E.S., Manktelow, B., Field, D.J. and James, D. (2003) Tables for predicting survival for preterm births are updated. *British Medical Journal* **327**:872

Draper, E.S., Manktelow, B.N., McCabe, C. and Field, D.J. (2004) The potential impact on costs and staffing of introducing clinical networks and British Association of Perinatal Medicine standards to the delivery of neonatal care. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F236-F240

Dubowitz, L.M.S., Dubowitz, V. and Goldberg, C. (1970) Clinical assessment of gestational age in newborn infants. *Journal of Pediatrics* **77**:1-10.

Duff, R.S. and Campbell, A.G.M. (1973) Moral and ethical dilemmas in the special care nursery. *New England Journal of Medicine* **289**:890-894.

Dyer, C. (1994) Inquiry into serial killer criticises hospital's response. *British Medical Journal* **308**:491

Eaton, D.G., Wertheim, D., Oozeer, R., Dubowitz, L.M. and Dubowitz, V. (1994) Reversible changes in cerebral activity associated with acidosis in preterm infants. *Acta Paediatrica* **83**:486-492.

Edwards, A.W.F. (1958) An analysis of Gressler's data on the human sex ratio. *Annals of Human Genetics* **23**:6-15.

Edwards, A.W.F. (1960) The meaning of binomial distribution. *Nature* **186**:1074

Effer, S.B., Mountquin, J.M., Farine, D., Saigal, S., Nimrod, C., Kelly, E. and Niyonsenga, T. (2002) Neonatal Survival Rates in 860 Singleton Live Births at 24 and 25 Weeks Gestational Age. A Canadian Multicentre Study. *BJOG - an International Journal of Obstetrics and Gynaecology* **109**:740-745.

Efron, B. and Tibshirani, R. (1993) An introduction to the bootstrap. New York, Chapman & Hall.

Einhorn, H.J. (1986) Accepting error to make less error. *Journal of Personality Assessment* **50**:387-395.

Elsmén, E., Hansen Pupp, I. and Hellström-Westas, L. (2004) Preterm male infants need more initial respiratory and circulatory support than female infants. *Acta Paediatrica* **93**:529-533.

Eriksen, P.S., Secher, N.J. and Weisbentzon, M. (1985) Normal growth of the fetal biparietal

diameter and the abdominal diameter in a longitudinal-study - an evaluation of the 2 parameters in predicting fetal weight. *Acta Obstetricia Et Gynecologica Scandinavica* **64**:65-70.

Ernest, J.M. (1998) Neonatal concequences of preterm PROM. *Clinical Obstetrics and Gynecology* **41**:827-831.

Evans, D.J. and Levene, M.I. (2001) Evidence of selection bias in preterm survival studies: a systematic review. *Archives of Disease in Childhood Fetal & Neonatal Edition* **84**, F79-F84

Expert Group on Learning from Adverse Events in the NHS (2000) An organisation with a memory. London, HMSO.

Farrell, T., Chien, P.F. and Gordon, A. (1999) Intrapartum umbilical artery Doppler velocimetry as a predictor of adverse perinatal outcome: a systematic review. *British Journal of Obstetrics and Gynaecology* **106**:783-792.

Fenton, A.C., Leslie, A. and Skeoch, C.H. (2004) Optimising neonatal transfer. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**, F215-F219.

Feyer, A-M. (2001) Fatigue: time to recognize and deal with an old problem. *British Medical Journal* **322**:808-809.

Field, D. and Draper, E.S. (1999) Survival and Place of Delivery Following Preterm Birth: 1994-96. *Archives of Disease in Childhood Fetal & Neonatal Edition* **80**, F111-F114

Field, D., Hodges, S., Mason, E. and Burton, P. (1991) Survival and place of treatment after premature delivery. *Archives of Disease in Childhood* **66**:408-411.

Field, D., Manktelow, B. and Draper, E.S. (2002) Bench marking and performance management in neonatal care: easier said than done! *Archives of Disease in Childhood Fetal & Neonatal Edition* **87**:F163-F164.

Field, D., Milligan, D., Skeoch, C. and Stephenson, T. (1997) Neonatal transport: time to change? *Archives of Disease in Childhood Fetal & Neonatal Edition* **76**, F1-F2

Field, D.J., Hodges, S., Mason, E., Burton, P., Yates, J. and Wale, S. (1989) The demand for neonatal intensive care. *British Medical Journal* **299**:1305-1308.

Field, D.J., Petersen, S., Clarke, M. and Draper, E.S. (2002) Extreme prematurity in the UK and Denmark: population differences in viability. *Archives of Disease in Childhood Fetal & Neonatal Edition* **87**, F172-F175

Fisher, E.S., Wennberg, J.E., Stukel, T.A. and Sharp, S.M. (1994) Hospital readmission rates for cohorts of Medicare beneficiaries in Boston and New Haven. *New England Journal of Medicine* **331** (15):989-995.

Fisher, L.D. (1996) Comments on Bayesian and Frequentist Analysis and Interpretation of Clinical Trials. *Controlled Clinical Trials* **17**:423-434.

Flanders, W.D., Shipp, C.C., FitzGerald, D.M. and Lin, L.S. (1994) Analysis of variation in mortality rates with small numbers. *Health Services Research* **29**:461-471.

Fleisher, B.E., Murthy, L., Lee, S., Constantinou, J.C., Benitz, W.E. and Stevenson, D.K. (1997) Neonatal severity of illness scoring systems: a comparison. *Clinical Pediatrics* **36** (4):223-227.

Fleisher, B.E., VandenBerg, K., Constantinou, J., Heller, C., Benitz, W.E., Johnson, A., Rosenthal, A. and Stevenson, D.K. (1995) Individualized development care for very-low-birth-weight premature infants. *Clinical Pediatrics* **34**:523-529.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003) Statistical Methods for Rates and Proportions. 3rd edn, New Jersey, Wiley.

Fowlie, P.W., Tarnow-Mordi, W.O., Gould, C.R. and Strang, D. (1998) Predicting outcome in very low birthweight infants using an objective measure of illness severity and cranial ultrasound scanning. *Archives of Disease in Childhood Fetal & Neonatal Edition* **78**:F175-F178

Freeman, J.V., Cole, T.J., Chinn, S., Jones, P.R.M., White, E.M. and Preece, M.A. (1995) Cross sectional stature and weight reference curves for the UK:1990. *Archives of Disease in Childhood* **73**:17-24.

Fretts, R.C., Schmittdiel, J., McLean, F.H., Usher, R.H. and Goldman, M.B. (1995) Increased maternal age and the risk of fetal death. *New England Journal of Medicine* **333**:953-957.

Fung, G., Bawden, K., Chow, P. and Yu, V. (2003) Chorioamnionitis and outcome in

extremely preterm infants. *Annals of the Academy of Medicine, Singapore* **32**:305-310.

Garcia, H., Villegas-Silva, R., Villanueva, D., Gonzalez-Cabello, H., Lopez-Padilla, M., Fajardo-Gutierrez, A., Martinez-Garcia, M. and Garduno-Espinosa, J. (2000) Validation of a prognostic index in the critically ill newborn. *La Revista de Investigacion Clinica* **52**:406-414.

Garite, T.J. and Freeman, R.K. (1982) Chorioamnionitis in the preterm gestation. *Obstetrics and Gynecology* **59**:539-545.

Gefland, A.E. (1995) Model determination using sample-based methods. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., (Eds.) *Markov Chain Monte Carlo in Practice*. London, Chapman & Hall, pp. 145-161.

Gelman, A (1995). Introducing Markov Chain Monte Carlo. In Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds) *Markov Chain Monte Carlo in Practice*. London, Chapman & Hall, pp. 131-143.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* **7**(4):457-472

Geyer, C.J. (1992) Practical Markov Chain Monte Carlo. *Statistical Science* **7**(4):457-511

Ghidini, A. and Spong, C.Y. (2001) Severe meconium aspiration syndrome is not caused by aspiration of meconium. *American Journal of Obstetrics and Gynecology* **185**:931-938.

Gibbs, J.L., Cunningham, D., de Leval, M., Monro, J. and Keogh, B. (2002) Paediatric cardiac hospital episode statistics are unreliable. *British Medical Journal* **330**:43-44.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., Mcneil, A.J., Sharples, L.D. and Kirby, A.J. (1993) Modeling complexity - applications of gibbs sampling in medicine. *Journal of the Royal Statistical Society Series B - Methodological* **55**:39-52.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1995). Introducing Markov Chain Monte Carlo. In Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds) *Markov Chain Monte Carlo in Practice*. London, Chapman & Hall, pp. 1-19.

Giuffrida, A., Gravelle, H. and Roland, M. (1999) Measuring quality of care with routine data: avoiding confusion between performance indicators and health outcomes. *British*

Medical Journal **319**:94-98.

Gjessing, H.K., Skjaerven, R. and Wilcox, A.J. (1999) Errors in gestational age: evidence of bleeding early in pregnancy. *American Journal of Public Health* **89**:213-218.

Glance, L.G., Osler, T. and Shinozaki, T. (2000) Effect of varying the case mix on the standardized mortality ratio and W statistic: a simulation study. *Chest* **117**:1112-1117.

Glance, L.G. and Osler, T.M. (2001) Comparing outcomes of coronary artery bypass surgery: is the New York Cardiac Surgery Reporting System model sensitive to changes in case mix? *Critical Care Medicine* **29**:2090-2096.

Glance, L.G., Osler, T.M. and Dick, A. (2002) Rating the quality of intensive care units: Is it a function of the intensive care unit scoring system? *Critical Care Medicine* **30**:1976-1982.

Goldenberg, R.L. (2002) The management of preterm labour. *Obstetrics and Gynecology* **100**:1020-1037.

Goldenberg, R.L., Hauth, J.C. and Andrews, W.W. (2000) Mechanisms of disease: intrauterine infection and preterm delivery. *New England Journal of Medicine* **342**:1500-1507.

Goldstein, H. (1995) Multilevel Statistical Models. 2nd edn, London, Arnold.

Goldstein, H. and Spiegelhalter, D.J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal Of The Royal Statistical Society Series A - Statistics In Society* **159**(Pt3):385-409.

Good, P.I. (1999) Resampling methods: a practical guide to data analysis. New York, Birkhäuser.

Gooi, A., Oei, J. and Lui, K. (2003) Attitudes of Level II obstetricians towards the care of the extremely premature infants: a national survey. *Journal of Paediatrics and Child Health* **39**:451-455.

Grant, A. and Glazener, C.M.A. (2003) Elective caesarean section versus expectant management for delivery of the small baby (Cochrane Review). In: *The Cochrane Library, Issue 3*, Oxford: Update Software

Gray, C.H., Howorth, P.J.N. and Rinsler, M.G. (1985) Clinical chemical pathology. London, Edward Arnold.

Gray, J.E. and Goldmann, D.A. (2004) Medication errors in the neonatal intensive care unit: special patients, unique issues. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F472-F473

Gray, J.E., Richardson, D.K., McCormick, M.C., Workman-Daniels, K. and Goldmann, D.A. (1992) Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. *Pediatrics* **90** (4):561-567.

Green, J. and Wintfeld, N. (1995) Report Cards on Cardiac-Surgeons: Assessing New-York States Approach. *New England Journal of Medicine* **332**:1229-1232.

Greisen, G. (2004) Meaningful care for babies born after 22, 23 or 24 weeks. *Acta Paediatrica* **93**:153-156.

Gruenwald, P. (1966) Growth of the human fetus. 1 Normal growth and its variation. *American Journal of Obstetrics and Gynecology* **94**:1112-1119.

Guildea, Z.E.S., Fone, D.L., Dunstan, F.D., Sibert, J.R. and Cartlidge, P.H.T. (2001) Social Deprivation and the Causes of Stillbirth and Infant Mortality. *Archives of Disease in Childhood* **84**:307-310.

Hadlock, F.P., Deter, R.L., Harrist, R.B. and Park, S.K. (1982) Fetal biparietal diameter: a critical re-evaluation of the relation to menstrual age by means of real-time ultasound. *Journal of Ultrasound in Medicine* **1**:97-104.

Harrell, F.E.Jr. (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York, Springer-Verlag.

Harrell, F.E.Jr., Lee, K.L. and Mark, D.E. (1996) Multivariable prognostic models:issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**:361-387.

Healthcare Quality and Analysis Division (2002) Report on Heart Attack Outcomes in California 1996-1998, Volume 2: Technical Guide. Sacramento, CA, California Office of Statewide Health Planning and Development.

Heckman, J.J. and Hotz, V.J. (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs. *Journal of the American Statistical Association* 84(408):862-880.

Hennekens, C.H. and Buring, J.E. (1987) Epidemiology in Medicine. Boston, Little, Brown and Company.

Henriksen, T.B., Wilcox, A.J., Hedegaard, M. and Secher, N.J. (1995) Bias in studies of preterm and postterm delivery due to ultrasound assessment of gestational age. *Epidemiology* **6**:533-537.

Hentschel, J., Friedel, G., Maier, R.F., Bassir, C. and Obladen, M. (1998) Predicting chronic lung disease in very low birthweight infants: comparison of 3 scores. *Journal of Perinatal Medicine* **26**:378-383.

Heydtmann, M. (2002) The nature of truth: Simpson's Paradox and the limits of statistical data. *QJM - an International Journal of Medicine* **95**:247-249.

Hofmeyr, G.J. and Kulier, R. (2003) Operative versus conservative management for 'fetal distress' in labour (Cochrane Review). In: *The Cochrane Library*, Issue 3. Oxford: Update Software.

Hollis, S., Yates, D.W., Woodford, M. and Foster, P. (1995) Standardized comparison of performance indicators in trauma: a new approach to case-mix variation. *Journal of Trauma* **38** (5):763-766.

Horbar, J.D., Badger, G.J., Carpenter, J.H., Fanaroff, A.A., Kilpatrick, S., Lacorte, M., Phibbs, R. and Soll, R.F. (2002) Trends in mortality and morbidity for very low birth weight infants:1991-1999. *Pediatrics* **110**:143-151.

Horbar, J.D., Badger, G.J., Lewit, E.M., Rogowski, J. and Shiono, P.H. (1997) Hospital and patient characteristics associated with variation in 28-day mortality rates for very low birth weight infants. *Pediatrics* **99**:149-156.

Horbar, J.D., Mcauliffe, T.L., Adler, S.M., Albersheim, S., Cassady, G., Edwards, W., Jones, R., Kattwinkel, J., Kraybill, E.N., Krishnan, V., Raschko, P. and Wilkinson, A.R. (1988) Variability in 28-day outcomes for very low birth-weight infants - an analysis of 11 neonatal intensive-care units. *Pediatrics* **82**:554-559.

Horbar, J.D., Onstad, L., Wright, E., Yaffe, S.J., Catz, C., Wright, L.L., Malloy, M.H., Rhoades, G.G., Wright, E., Gordon, T., Onstad, L., Phillips, E., Oh, W., Cassady, G., Philips, J., Lucey, J.F., Horbar, J.D., Fanaroff, A.A., Hack, M., Tyson, J.E., Uauy, R., Poland, R., Shankaran, S., Little, G., Edwards, W., Korones, S.B., Cooke, R., Bauer, C.R. and Bandstra, E.S. (1993a) Predicting mortality risk for infants weighing 501 to 1500 grams at birth - a National Institutes of Health Neonatal Research Network Report. *Critical Care Medicine* **21**:12-18.

Horbar, J.D., Wright, E.C., Onstad, L., Philips, J.B., Cassady, G., Fanaroff, A.A., Hack, M., Edwards, W., Little, G.A., Foley, K.A., Bauer, C.R., Bandstra, E.S., Yaffe, S.J., Wright, L.L., Malloy, M.H., Korones, S.B., Cooke, R., Tyson, J.E., Uauy, R.D., Lucey, J.F., Shankaran, S. and Ostrea, E. (1993b) Decreasing mortality associated with the introduction of surfactant therapy: an observational study of neonates weighing 601 to 1300 grams at birth. *Pediatrics* **92**:191-196.

Hosmer, D.W., Hosmer, T., Lecessie, S. and Lemeshow, S. (1997) A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* **16**:965-980.

Hosmer, D.W. and Lemeshow, S. (1980) A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics* **A10**:1043-1069.

Hosmer, D.W. and Lemeshow, S. (1995) Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine* **14** (19):2161-2172.

Hosmer, D.W. and Lemeshow, S. (1997) Confidence interval estimates of an index of quality performance based on logistic regression models (Reply). *Statistics in Medicine* **16**:1303

Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd edn, New York, Whiley.

Howell, J. (1995) Standardised mortality ratios. *Lancet* **346**:904.

Howell, J. (2002) Performance league tables - league tables are unreasonably simple. *British Medical Journal* **324**:542.

Huang, I.-C., Frangakis, C., Dominici, F., Diette, G.B. and Wu, A.W. (2005) Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* 40(1):253-278.

Hubbard, M. and Haines, L. (2004) A national programme of neonatal audit: report of a feasibility study. London, Neonatal Clinical Audit Programme.

Iezzoni, L.I. (1994) Using risk-adjusted outcomes to assess clinical practice: an overview of issues pertaining to risk adjustment. *Annals of Thoracic Surgery* **58** (6):1822-1826.

Iezzoni, L.I. (1996) 100 apples divided by 15 red herrings: a cautionary tale from the mid-19th century on comparing hospital mortality rates. *Annals of Internal Medicine* **124** (12):1079-1085.

Iezzoni, L.I. (1997) The risks of risk adjustment. *Journal of the American Medical Association* **278**(19):1600-1607.

Iezzoni, L.I., Ash, A.S., Coffman, G.A. and Moskowitz, M.A. (1992) Predicting in-hospital mortality. A comparison of severity measurement approaches. *Medical Care* **30**(4):347-359.

Iezzoni, L.I., Ash, A.S., Shwartz, M., Daley, J., Hughes, J.S. and Mackiernan, Y.D. (1995) Predicting who dies depends on how severity is measured: implications for evaluating patient outcomes. *Annals of Internal Medicine* **123**(10):763-770.

Iezzoni, L.I., Ash, A.S., Shwartz, M., Daley, J., Hughes, J.S. and Mackiernan, Y.D. (1996a) Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method. *American Journal of Public Health* **86**(10):1379-1387.

Iezzoni, L.I., Shwartz, M., Ash, A.S. and Mackiernan, Y.D. (1996b) Predicting in-hospital mortality for stroke patients: results differ across severity-measurement methods. *Medical Decision Making* **16**(4):348-356.

Ingemarsson, I. (2003) Gender aspects of preterm birth. *BJOG: an International Journal of Obstetrics and Gynaecology* **110**(Suppl 20):34-38.

International Neonatal Network, S.N.C.N.C.S.G. (2000) Risk adjusted and population based studies of the outcome for high risk infants in Scotland and Australia. *Archives of Disease in Childhood Fetal & Neonatal Edition* **82**:F118-F123

Italian Collaborative Group on Preterm Delivery (1988) Prenatal and postnatal factors affecting short-term survival of very low birth weight infants. *European Journal of Pediatrics* **147**:468-471.

Ivanov, J., Tu, J.V. and Naylor, D. (1999) Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation* **99**:2098-2104.

Iyasu, S., Tomashek, K. and Barfield, W. (2002) Infant mortality and low birth weight among black and white infants - United States 1980-2000. *Journal of the American Medical Association* **288**:825-826.

Jain, A. and Fleming (2004) Project 27/28. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F14-F16

James, D.S. and Leadbeatter, S. (1997) Detecting homicide in hospital. *Journal of the Royal College of Physicians of London* **31**:296-298.

Jankowski, R. (1999) What do hospital admission rates say about primary care? *British Medical Journal* **319**(7202):67-68.

Jarman, B. (1983) Identification of underprivileged areas. *British Medical Journal* **286**:1705-1709.

Jarman, B., Gault, S., Alves, B., Hider, A., Dolan, S., Cook, A., Hurwitz, B. and Iezzoni, L.I. (1999) Explaining differences in English hospital death rates using routinely collected data. *British Medical Journal* **318**:1515-1520.

Jazayeri, A., Politz, L., Tsibris, J.C.M., Queen, T. and Spellacy, W.N. (2000) Fetal erythropoitin levels in pregnancies complicated by meconium passage: does meconium suggest fetal hypoxia? *American Journal of Obstetrics and Gynecology* **183**:188-190.

Jencks, S.F., Daley, J., Draper, D., Thomas, N., Lenhart, G. and Walker, J. (1988) Interpreting hospital mortality data. The role of clinical risk adjustment. *Journal of the American Medical Association* **260**(24):3611-3616.

Jenkins, K., Gauvreau, K., Newburger, J., Spray, T., Moller, J. and Iezzoni, L. (2002) Consensus-based method for risk-adjustment for surgery for congenital hearth disease. *Journal of Thoracic and Cardiovascular Surgery* **123**:110-118.

Jenkins, K.J. and Gauvreau, K. (2002) Centre-specific differences in mortality:preliminary analyses using the Risk Adjustment in Congenital Heart Surgery (RACHS-1) method. *Journal*

of Thoracic and Cardiovascular Surgery **124**:97-104.

Jenkins, K.J., Newburger, J.W., Lock, J.E., Davis, R.B., Coffman, G.A. and Iezzoni, L.I. (1995) In-hospital mortality for surgical repair of congenital heart defects: preliminary observations of variation by hospital caseload. *Pediatrics* **95**(3):323-30.

Jiang, J. (2001) Goodness-of-fit tests for mixed model diagnostics. *Annals of Statistics* **29**(4):1137-1164.

Jolly, M., Sebire, N., Harris, J., Robinson, S. and Regan, L. (2000) The risks associated with pregnancy in women aged 35 years or older. *Human Reproduction* **15**:2433-2437.

Jones, J.M., Redmond, A.D. and Templeton, J. (1995) Uses and abuses of statistical models for evaluating trauma care. *Journal of Trauma* **38**:89-93.

Joseph, K.S., Kramer, M.S., Allen, A.C., Mery, L.S., Platt, R.W. and Wen, S.W. (2001) Implausible birth weight for gestational age. *American Journal of Epidemiology* **153**:110-113.

Joyce, R., Webb, R. and Peacock, J.L. (2004) Associations between perinatal interventions and hospital stillbirth rates and neonatal mortality. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**(1):F51-F56.

Joyce, R., Webb, R., Peacock, J.L. and Stirland, H. (2002) Adjusted Mortality Rates: a Tool for Creating More Meaningful League Tables for Stillbirth and Infant Mortality Rates. *Public Health* **116**:315-321.

Julious, S.A., Nicholl, J. and George, S. (2001) Why do we continue to use standardized mortality ratios for small area comparisons? *Journal of Public Health Medicine* **23**:40-46.

Kahn, D.J., Richardson, D.K., Gray, J.E., Bednarek, F., Rubin, L.P., Shah, B., Frantz, I.D. 3rd and Pursley, D.M. (1998) Variation among neonatal intensive care units in narcotic administration. *Archives of Pediatrics & Adolescent Medicine* **152**(9):844-51.

Kahn, K.L., Brook, R.H., Draper, D., Keeler, E.B., Rubenstein, L.V., Rogers, W.H. and Kosecoff, J. (1988) Interpreting hospital mortality data. How can we proceed? *Journal of the American Medical Association* **260**(24):3625-8.

Kallen, B. (1995) A birth-weight for gestational-age standard based on data in the Swedish

Medical Birth Registry:1985-1989. *European Journal of Epidemiology* **11**:601-606.

Kassirer, J.P. (1994) The use and abuse of practice profiles. *New England Journal of Medicine* **330**(9):634-6.

Kaushal, R., Bates, D.W., Landrigan, C., McKenna, K.J., Clapp, M.D., Federico, F. and Goldmann, D.A. (2001) Medication errors and adverse drug events in pediatric inpatients. *Journal of the American Medical Association* **285**:2114-2120.

Keene, A.R. and Cullen, D.J. (1983) Therapeutic Intervention Scoring System - Update 1983. *Critical Care Medicine* **11**:1-3.

Keiding, N. (1985) Standardized mortality ratio and statistical analysis: historical perspective. *Biometrics* **41**:1096.

Keogh, B., Spiegelhalter, D., Bailey, A., Roxburgh, J., Magee, P. and Hilton, C. (2004) The legacy of Bristol: public disclosure of individual surgeons' results. *British Medical Journal* **329**:450-454.

Kiely, J.L. (1998) What is the population-based risk of preterm birth among twins and other multiples? *Clinical Obstetrics and Gynecology* **41**:3-11.

Kinnell, H.G. (2000) Serial homicide by doctors: Shipman in perspective. *British Medical Journal* **321**:1594-1596.

Knaus, W.A., Wagner, D.P., Zimmerman, J.E. and Draper, E.A. (1993) Variations in mortality and length of stay in intensive-care units. *Annals of Internal Medicine* **118**:753-761.

Kohn, K.T., Corrigan, J.M., Donaldson, M.S. (2000) To Err is Human: Building a Safer Health Care System. Washington DC, National Academy Press.

Kotz, S. and Johnson, N.L. (1982) Encyclopedia of Statistical Sciences: I. New York, Wiley.

Kramer, M.S., Goulet, L., Lydon, J., Seguin, L., McNamara, H., Dassa, C., Platt, R.W., Chen, M.F., Gauthier, H., Genest, J., Kahn, S., Libman, M., Rozen, R., Masse, A., Miner, L., Asselin, G., Benjamin, A., Klein, J. and Koren, G. (2001) Socio-economic disparities in preterm birth: causal pathways and mechanisms. *Paediatric & Perinatal Epidemiology* **15**(Suppl 2):104-23.

Kramer, M.S., McLean, F.H., Boyd, M.E. and Usher, R.H. (1988) The validity of gestational age estimation by menstrual dating in term, preterm, and postterm gestations. *Journal of the American Medical Association* **260**:3306-3308.

Kramer, M.S., Platt, R., Yuan, H., McNamara, H. and Usher, R.H. (1999) Are all growth-restricted newborns created equal(ly)? *Pediatrics* **103**:599-602.

Kramer, M.S., Platt, R.W., Wen, S.W., Joseph, K.S., Allen, A., Abrahamowicz, M., Blondel, B. and Breart, G. (2001) A new and improved population-based Canadian reference for birth weight for gestational age. *Pediatrics* **108**(2):E35

Källén, K. (2002) Mid-trimester ultrasound prediction of gestational age: advantages and systematic errors. *Ultrasound in Obstetrics & Gynecology* **20**:558-563.

Lackman, F., Capewell, V., Richardson, B., daSilva, O. and Gagnon, R. (2001) The risks of spontaneous preterm delivery and perinatal mortality in relation to size at birth according ot fetal versus neonatal growth standards. *American Journal of Obstetrics and Gynecology* **184**:946-953.

Lal, K.M., Manktelow, B.N., Draper, E.S. and Field, D.J. (2003) Chronic lung disease of prematurity and intrauterine growth retardation: a population-based study. *Pediatrics* **111**:483-487.

Lambert, P.C., Sutton, A.J., Abrams, K.R. and Jones, D.R. (2002) A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology* 55(1):86-94.

Landon, B., Iezzoni, L.I., Ash, A.S., Shwartz, M., Daley, J., Hughes, J.S. and Mackiernan, Y.D. (1996) Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry* **33**(2):155-66.

Lange, N. and Ryan, L. (1989) Assessing normality in random effects models. *Annals of Statistics* **17**:624-642.

Langford, I.H. (1997) Statistical basis of public policy. Bayesian analysis should be used instead of league tables of performance. *British Medical Journal* **314**(7073):73-4.

Larroque, B., Breart, G., Kaminski, M., Dehan, M., Andre, M., Burguet, A., Grandjean, H.,

Ledesert, B., Leveque, C., Maillard, F., Matis, J., Roze, J.C. and Truffert, P. (2004) Survival of very preterm infants: Epipage, a population based cohort study. *Archives of Disease in Childhood* **89**:139-144.

Leary, T., Ridley, S., Burchett, K., Kong, A., Chrispin, P. and Wright, M. (2002) Assessing critical care unit performance: a global measure using graphical analysis. *Anaesthesia* **57**:751-755.

Lee, S.K., Mcmillan, D.D., Ohlsson, A., Pendray, M., Synnes, A., Whyte, R., Chien, L.Y. and Sale, J. (2000) Variations in practice and outcomes in the Canadian NICU Network: 1996-1997. *Pediatrics* **106**:1070-1079.

Lee, S.K., Zupancic, J.A.F., Pendray, M., Thiessen, P., Schmidt, B., Whyte, R., Shorten, D., Stewart, S. and The Canadian Neonatal Network (2001) Transport Risk Index of Physiologic Stability: a practical system for assessing infant transport care. *Journal of Pediatrics* **139**:220-226.

Lee, W.C. (1999) Properties of the Geometrically Averaged Ratio, an alternative standardized measure. *Epidemilogy* **10**:456-459.

Lee, W.C. (2002) Standardization using the harmonically weighted ratios: internal and external comparisons. *Statistics in Medicine* **21**:247-261.

Lemeshow, S., Klar, J. and Teres, D. (1995) Outcome prediction for individual intensive-care patients: useful, misused, or abused. *Intensive Care Medicine* **21**:770-776.

Leon, D.A. (1991) Influence of birth weight on differences in infant mortality by social class and legitimacy. *British Medical Journal* **303**:964-967.

Letko, M.D. (1996) Understanding the Apgar score. *Journal of Obstetric, Gynecologic, and Neonatal Nursing* **25**:299-303.

Levene, M. (2004) Is intensive care for very immature babies justified? *Acta Paediatrica* **93**:149-152.

Leyland, A.H. and Boddy, F.A. (1998) League tables and acute myocardial infarction. *Lancet* **351**(9102):555-558.

Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**:13-22.

Lie, R.T. (2000) Intersecting perinatal mortality curves by gestational age: are appearances deceiving? *American Journal of Epidemiology* **152**:1117-1119.

Liggins, G.C. and Howie, R.N. (1972) A controlled trial of antepartum glucocorticoid treatment for prevention of the respiratory distress syndrome in premature infants. *Pediatrics* **50**:515-525.

Lindley, D.V. (2005) Foundations of Probability. In: Armitage, P. and Colton, T., (Eds.), *Encyclopedia of Biostatistics*. Chichester, Wiley: pp.1993-2001.

Lindsey, J (2005). repeated: Non-normal Repeated Measurements Models. R package version 1.0. www.luc.ac.be/~jlindsey/rcode.html

Liu, S., Joseph, K.S., Kramer, M.S., Allen, A.C., Suave, R., Rusen, I.D. and Wen, S.W. (2002) Relationship of prenatal diagnosis and pregnancy termination to overall infant mortality in Canada. *Journal of the American Medical Association* **287**:1561-1567.

Livingston, J. (1990) Interrater reliability of the Apgar score in term and premature infants. *Applied Nursing Research* **3**:164-165.

Livingston, J.C., Livingston, L.W., Ramsey, R. and Sibai, B.M. (2004) Second-trimester asychronous multifetal delivery results in poor perinatal outcome. *Obstetrics and Gynecology* **103**:77-81.

Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C. and Gallivan, S. (1997) Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* **350** (9085):1128-1130.

Low, J.A., Panagiotopoulos, C. and Derrick, E.J. (1994) Newborn complications after intrapartum asphyxia with metabolic acidosis in the term fetus. *American Journal of Obstetrics and Gynecology* **170**:1081-1087.

Low, J.A., Panagiotopoulos, C. and Derrick, E.J. (1995a) Fetus-placenta-newborn: newborn complications after intrapartum asphyxia with metabolic acidosis in the preterm fetus. *American Journal of Obstetrics and Gynecology* **172**:805-810.

Low, J.A., Simpson, L.L., Tonni, G. and Chamberlain, S. (1995b) Fetus-placenta-newborn: limitations in the clinical prediction of intrapartum fetal asphyxia. *American Journal of Obstetrics and Gynecology* **172**:801-804.

Luft, H.S. and Brown, B.W. (1993) Calculating the probability of rare events: why settle for an approximation. *Health Services Research* **28**:419-439.

Lyon, A.J. and Stenson, B. (2004) Cold comfort for babies. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**(1):F93

Macfarlane, A. and Mugford, M. (2000) Birth counts: statistics of pregnancy & childbirth. Volume 1. 2nd edn, Norwich, The Stationery Office.

Manchester Heart Centre (2005)
Avaliable at http://www.manchesterheartcentre.org/common/datamenu.php

Maier, R.F., Rey, M., Metze, B.C. and Obladen, M. (1997) Comparison of mortality risk: a score for very low birthweight infants. *Archives of Disease in Childhood Fetal & Neonatal Edition* **76**(3):F146-F150.

Manktelow, B.N. and Draper, E.S. (2003) Neonatal mortality and socio-economic status: a case-control study. *Gaceta Sanitaria* **17**:69

Manktelow, B.N., Draper, E.S., Annamalai, S. and Field, D. (2001) Factors affecting the incidence of chronic lung disease of prematurity in 1987:1992, and 1997. *Archives of Disease in Childhood Fetal & Neonatal Edition* **85**:F33-F35

Mant, J. and Hicks, N. (1995) Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *British Medical Journal* **311**(7008):793-796.

Margetts, B.M., Yusof, S.M., Dallal, Z.A. and Jackson, A.A. (2002) Persistence of lower birth weight in second generation South Asian babies born in the United Kingdom. *Journal of Epidemiology and Community Health* **56**:684-687.

Marlow, N. (2002) Illness serverity measures in neonatal intensive care. *Acta Paediatrica* **91**:367-368.

Marlow, N. (2004) Neurocognitive outcome after very preterm birth. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F224-F228

Marshall, E.C. and Spiegelhalter, D.J. (1998a) Comparing institutional performance using Markov chain Monte Carlo methods. In: Everitt, B.S. and Dunn, G., (Eds.) S*tatistical analysis of medical data: new developments*, London, Arnold, pp. 229-249

Marshall, E.C. and Spiegelhalter, D.J. (1998b) Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* **316** (7146):1701-1704.

Martuzzi, M. and Elliott, P. (1998) Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. *Annals of Epidemiology* **8**:52-55.

Martuzzi, M., Grundy, C. and Elliott, P. (1998) Perinatal mortality in an English Health Region: geographical distribution and association with socio-economic factors. *Paediatric and Perinatal Epidemiology* **12**:263-276.

McCormick, M.C. (1997) The outcomes of very low birth weight infants: are we asking the right questions? *Pediatrics* **99**:869-876.

McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2 edn, London, Chapman & Hall.

McIntire, D.D., Bloom, S.L., Casey, B.M. and Leveno, K.J. (1999) Birth weight in relation to morbidity and mortality among newborn infants. *New England Journal of Medicine* **340**:1234-1238.

McNutt, L., Wu, C., Xue, X. and Hafner, J.P. (2003) Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology* **157**:940-943.

Mc Nutt, R.A., Abrams, R., Aron, D.C., for the Patient Safety Committee (2002) Patient safety efforts should focus on medical errors. *Journal of the American Medical Association* **287**(15):1997-2001.

Mead, M. (1996) The diagnosis of foetal distress: a challenge to midwives. *Journal of Advanced Nursing* **23**:975-983.

Meis, P.J., Michielutte, R., Peters, T.J., Wells, H.B., Sands, R.E., Coles, E.C. and Johns, K.A. (1995) Factors associated with preterm birth in Cardiff, Wales. *American Journal of Obstetrics and Gynecology* **173**:590-596.

Menke, J.A., Broner, C.W., Campbell, D.Y., McKissick, M.Y. and Edwards-Beckett, J.A. (2001) Computerized clinical documentation system in the pediatric intensive care unit . *BMC Medical Informatics and Decision Making* **1**:3

Michel, P., Quenon, J.L., de Sarasqueta, A.M. and Scemama, O. (2004) Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. *British Medical Journal* **328**:199-203.

Milburn, A. (2002) HC Deb (2001-2002) 378, col 448.

Miller, F.C., Sacks, D.A., Yeh, S.-Y., Paul, R.H., Schifrin, B.S., Martin, C.B. and Hon, E.H. (1975) Significance of meconium during labor. *American Journal of Obstetrics and Gynecology* **122**:573-580.

Miller, M.E., Hui, S.L. and Tierney, W.M. (1991) Validation techniques for logistic regression models. *Statistics in Medicine* **10**:1213-1226.

Miller, M.E., Langefeld, C.D., Tierney, W.M., Hui, S.L. and McDonald, C.J. (1993) Validation of probabilistic predictions. *Medical Decision Making* **13**:49-53.

Milner, R.D.G. and Richards, B. (1974) An analysis of birth weight by gestational age of infants bornin England and Wales:1967 to 1971. *The Journal of Obstetrics and Gynaecology of the British Empire* **81**:956-967.

Mohammed, M.A., Cheng, K.K., Rouse, A. and Marshall, T. (2001a) Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* **357**:463-467.

Mohammed, M.A., Cheng, K.K., Rouse, A., Marshall, T. and Duffy, J. (2001b) Was Bristol an outlier? *Lancet* **358**:2083-2084.

Molloy, E.J., O'Neill, A.J., Grantham, J.J., Sheridan-Pereira, M., Fitzpatrick, J.M., Webb, D.W. and Watson, R.W.G. (2004) Labor promoted neonatal neutrophil survival and lipopolysaccharide responsiveness. *Pediatric Research* **56**:99-103.

Morgan, B.L., Chao, C.R., Iyer, V. and Ross, M.G. (2002) Correlation between fetal scalp blood samples and intravascular blood pH, pO2 and oxygen saturation measurements. *Journal of Maternal, Fetal and Neonatal Medicine* **11**:325-328.

Morley, C.J., Thornton, A.J., Cole, T.J., Hewson, P.H. and Fowler, M.A. (1991) Baby Check: a scoring system to grade the severity of acute systemic illness in babies under 6 months old. *Archives of Disease in Childhood* **66**:100-105.

Morrison, J.J. and Rennie, J.M. (1997) Clinical, scientific and ethical aspects of fetal and neonatal care at extremely preterm periods of gestation. *British Journal of Obstetrics and Gynaecology* **104**:1341-1350.

Motulsky, H.J. (1995) Intuitive Biostatistics. Oxford, Oxford University Press.

Mulaik, S.A., Raju, N.S. and Harshman, R.A. (1997) There is a time and a place for significance testing. In: Harlow, L.L., Mulaik, S.A. and Steiger, J.H., (Eds.) *What if there were no significance tests?* New Jersey, Lawrence Erldaum Associates

Mustafa, G. and David, R.J. (2001) Comparative accuracy of clinical estimate versus menstrual gestational age in computerized birth certificates. *Public Health Reports* **116**:15-21.

Nagel, H.T.C., Vandenbussche, F.P.H.A., Oepkes, D., Jennekens-Schinkel, A., Laan, L.A.E.M. and Gravenhorst, J.B. (1995) Follow-up of children born with an umbilical arterial blood pH < 7. *American Journal of Obstetrics and Gynecology* **173**:1758-1764.

National Collaborating Centre for Women's and Children's Heath (2003) Antenatal care: routine care for the healthy pregnant woman. London, RCOG Press.

National Institute for Clinical Excellence (2004) Caesarean Section: Guideline 13. London, National Institute for Clinical Excellence.

Naylor, C., Vanderhal, A., Hoble, C., Forbis, S. and Sola, A. (2001) Caesarean delivery for extremely low birth weight infants 500-750g in breech presentation: what are the benefits? *American Journal of Obstetrics and Gynecology* **184**:S194

Neilson, J.P. (2003) Ultrasound for fetal assessment in early pregnancy (Cochrane Review). In: *The Cochrane Library, Issue 3*, Oxford, Update Software

Neison, F.G.P. (1844) On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census. *Journal of the Statistical Society of London* **7**:40-68.

Nelson, L.H., Anderson, R.L., O'Shea, T.M. and Swain, M. (1994) Expectant management of preterm premature rupture of the membranes. *American Journal of Obstetrics and Gynecology* **171**:350-258.

Neuhaus, J.M., Hauck, W.W. and Kalbfleisch, J.D. (1992) The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**(4):755-762.

New York State Department of Health (1998a) Coronary artery bypass surgery in New York State 1994-96. New York, New York State Department of Health.

New York State Department of Health (1998b) Press Release - New York State Department of Health. New York, New York State Department of Health.

New York State Department of Health (2000) Coronary artery bypass surgery in New York State 1995-1997. New York, New York State Department of Health.

New York State Department of Health (2004) Adult cardiac surgery in New York State 1998-2000. New York, New York State Department of Health.

NHS Executive (1998) The new NHS: a national framework for assessing performance. Leeds, NHSE.

Nightingale, F. (1863) Notes on Hospitals. 3rd edn, London, Longman, Roberts, and Green.

Normand, S.L.T., Glickman, M.E. and Gatsonis, C.A. (1997) Statistical methods for profiling providers of medical care: issues and applications. *Journal Of The American Statistical Association* **92** (439):803-814.

Nutley, S. and Smith, P.C. (1998) League tables for performance improvement in health care. *Journal of Health Services & Research Policy* **3** (1):50-57.

O'Connor, G.T., Plume, S.K., Olmstead, E.M., Morton, J.R., Maloney, C.T., Nugent, W.C., Hernandez, F., Clough, R., Leavitt, B.J., Coffin, L.H., Marrin, C.A.S., Wennberg, D.,

Birkmeyer, J.D., Charlesworth, D.C., Malenka, D.J., Quinton, H.B. and Kasper, J.F. (1996) A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. *Journal of the American Medical Association* **275**:841-846.

Oei, J., Askie, L.M., Tobiansky, R. and Lui, K. (2000) Attitudes of neonatal clinicians towards resuscitation of the extremely premature infant: a exploratory survey. *Journal of Paediatrics and Child Health* **36**:357-362.

Office of National Statistics. (2003a)
Available at http://www.statistics.gov.uk/STATBASE/ Expodata/Spreadsheets/D7149.xls.

Office of National Statistics (2003b) Series DH3 No. 34. Mortality Statistics:childhood, infant and perinatal. London, HMSO.

Office of the Deputy Prime Minister (2003) The English Indices of Deprivation 2004. London, ODPM.

Ogle, W. (1886) Suicides in England and Wales in relation to age, sex, season, and occupation. *Journal of the Statistical Society of London* **49**:101-135.

Oja, H., Koiranen, M. and Rantakallio, P. (1991) Fitting mixture-models to birth-weight data: a case-study. *Biometrics* **47**:883-897.

Ott, W.J. (1993) Intrauterine growth retardation and preterm delivery. *American Journal of Obstetrics and Gynecology* **168**:1710-1717.

Paneth, N. (1992) Very-low-birth-weight: a problematic cohort for epidemiologic studies of very small or immature neonates. *American Journal of Epidemiology* **136**:767.

Paneth, N., Wallenstein, S., Kiely, J.L. and Susser, M. (1982) Social class indicators and mortality in low birth weight infants. *American Journal of Epidemiology* **116**:364-375.

Papworth Hospital (2005)
Available at http://www.papworthpeople.com/frame_clinical.asp?section=clinical

Parer, J.T. (2003) Electronic fetal heart rate monitoring: a story of survival. *Obstetrical and Gynecological Survey* **58**:561-563.

Parker, J.D. and Schoendorf, K.C. (2002) Implications of cleaning gestational age data.

Paediatric and Perinatal Epidemiology **16**:181-187.

Parmanum, J., Field, D., Rennie, J. and Steer, P. (2000) National census of availability of neonatal intensive care. *British Medical Journal* **321**:727-729.

Parmar, M.K.B., Griffiths, G.O., Spiegelhalter, D.J., Souhami, R.L., Altman, D.G. and Van Der Scheuren, E. (2001) Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *Lancet* **358**:375-381.

Parry, G., Tucker, J. and Tarnow-Mordi, W. (2003a) Volume of procedures and outcome of treatment. *British Medical Journal* **326**:280

Parry, G., Tucker, J. and Tarnow-Mordi, W. (2003b) CRIB II: an update of the Clinical Risk Index for Babies score. *Lancet* **361**:1789-1791.

Parry, G.J., Gould, C.R., McCabe, C.J. and Tarnow-Mordi, W.O. (1998) Annual league tables of mortality in neonatal intensive care units: longitudinal study. International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group. *British Medical Journal* **316** (7149):1931-1935.

Parsonnet, V., Dean, D. and Bernstein, A.D. (1989) A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation* **79**(6 Pt2):I3-12.

Peacock, J.L., Bland, J.M. and Anderson, R. (1995) Preterm delivery: effects of socioeconomic factors, psychological stress, smoking, alcohol and caffeine. *British Medical Journal* **311**:531-535.

Pearce, J.M. and Steel, S.A. (1987) A manual of labour ward practice. Chichester, Wiley.

Penn, Z. and Ghaem-Maghami, S. (2001) Indications for Caesarean Section. *Best Practice & Research in Clinical Obstetrics & Gynaecology* **15**:1-15.

Penn, Z. and Steer, P. (1990) Reasons for declining participation in a prospective randomised trial to determine the optimum mode of delivery of the preterm breech. **11**:226-231.

Pennsylvania Health Care Cost Containment Council (1992a) A consumer guide to coronary artery bypass graft surgery: Vol I. 1990 data. Harrisberg, Pennsylvania Health Care Cost

Containment Council.

Pennsylvania Health Care Cost Containment Council (1992b) Coronary artery bypass graft surgery: a technical report: Vol. I. 1990 data. Harrisberg, Pennsylvania Health Care Cost Containment Council.

Perneger, T.V. (1998) What's wrong with Bonferroni adjustments. *British Medical Journal* **316**:1236-1238

Peterson, E.D., Moore, D., Muhlbaier, L.H., Delong, E.R. and Grosswald, R. (1998) Predicting mortality following PTCA: results from NCN. *Journal Of The American College Of Cardiology* **31**(2 SA):A179.

Phibbs, C.S., Bronstein, J.M., Buxton, E. and Phibbs, R.H. (1996) The effects of patient volume and level of care at the hospital of birth on neonatal mortality. *Journal of the American Medical Association* **276**(13):1054-1059.

Pickersgill, T. (2001) The European working time directive for doctors in training. *British Medical Journal* **323**:1266.

Platt, R.W., Abrahamowicz, M., Kramer, M.S., Joseph, K.S., Mery, L., Blondel, B., Breart, G. and Wen, S.W. (2001) Detecting and eliminating erroneous gestational ages: a normal mixture model. *Statistics in Medicine* **20**:3491-3503.

Pollack, M.M., Koch, M.A., Bartel, D.A., Rapoport, I., Dhanireddy, R., El-Mohandes, A.A.E., Harkavy, K. and Subramanian, K.N.S. (2000) A comparison of neonatal mortality risk prediction models in very low birth weight infants. *Pediatrics* **105**:1051-1057.

Poloniecki, J. (1998) Half of all doctors are below average. *British Medical Journal* **316** (7146):1734-1736.

Poloniecki, J., Valencia, O. and Littlejohns, P. (1998) Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal* **316**:1697-1700.

Pregibon, D. (1981) Logistic regression diagnosis. *Annals of Statistics* **9**:705-724.

Pschirrer, E.R. and Monga, M. (2000) Risk factors for preterm labor. *Clinical Obstetrics and*

Gynecology **43**:727-734.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rabelink, I.A., Degen, J.E., Kessels, M.E., Nienhuis, S.J., Ruissen, C.J. and Hoogland, H.J. (1994) Variation in early fetal growth. *European Journal of Obstetrics, Gynecology & Reproductive Biology* **53**:39-43.

Rabilloud, M., Ecochard, R. and Esteve, J. (2001) Maternity hospitals ranking on prophylactic caesarean section rates: uncertainty associated with ranks. *European Journal of Obstetrics Gynecology and Reproductive Biology* **94**:139-144.

Ramin, S.M., Gilstrap III, L.C., Leveno, K.J., Burris, J. and Little, B.B. (1989) Umbilical artery acid-base status in the preterm infant. *Obstetrics and Gynecology* **74**:256-258.

Rao, J.N. (2001) Hospital league tables. *British Medical Journal* **322**:992

Rapoport, J., Teres, D., Lemeshow, S. and Gehlbach, S. (1994) A method for assessing the clinical performance and cost-effectiveness of intensive care units: a mulitcenter inception cohort study. *Critical Care Medicine* **22**:1385-1391.

Rautonen, J., Makela, A., Boyd, H., Apajasalo, M. and Pohjavuori, M. (1994) CRIB and SNAP: assessing the risk of death for preterm neonates. *Lancet* **343** (8908):1272-3.

RCOG & RCM (1999) Towards safer childbirth. Minimum standards for the organisation of labour wards. Report of a joint working party. London, RCOG Press.

Regev, R.H., Lusky, A., Dolfin, T., Litmanovitz, I., Arnon, S. and Reichman, B. (2003) Excess mortality and morbidity among small-for-gestational-age premature infants: a population-based study. *Journal of Pediatrics* **143**:186-191.

Reid, F.D.A., Cook, D.G. and Majeed, A. (1999) Explaining Variation in Hospital Admission Rates Between General Practices: Cross Sectional Study. *British Medical Journal* **319**:98-103.

Rennie, J.M. and Roberton, N.R.C. (2002) A manual of neonatal intensive care. Rennie, J.M.

and Roberton, N.R.C., (Eds.) 4th edn, London, Arnold.

Richardson, D.K., Corcoran, J.D., Escobar, G.J. and Lee, S.K. (2001) SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *Journal of Pediatrics* **138**:92-100.

Richardson, D.K., Gray, J.E., McCormick, M.C., Workman, K. and Goldmann, D.A. (1993) Score for Neonatal Acute Physiology: a physiologic severity index for neonatal intensive care. *Pediatrics* **91** (3):617-23.

Richardson, D.K., McCormick, M.C., Gray, J.E. and Goldmann, D.A. (1994) CRIB and SNAP. *Lancet* **344**(8915):124-5.

Richardson, D.K., Phibbs, C.S., Gray, J.E., McCormick, M.C., Workman-Daniels, K. and Goldmann, D.A. (1993) Birth weight and illness severity: independent predictors of neonatal mortality. *Pediatrics* **91**(5):969-75.

Ridley, S.A. (2002) Uncertainty and scoring systems. *Anaesthesia* **57**:761-767.

Rijken, M., Stoelhorst, G.M.S.J., Martens, S.E., van Zwieten, P.H.T., Brand, R., Wit, J.M. and Veen, S. (2003) Mortality and neurologic, mental, and psychomotor development at 2 years in infants born less than 27 weeks' gestation: the Leiden Follow-Up Project on Prematurity. *Pediatrics* **112**:351-358.

Ritz, C. (2004) Goodness-of-fit tests for mixed models. *Scandinavian Journal of Statistics* **31**:433-458.

Rixom, A. (2002) Performance league tables - use of indirect standardisation is inappropriate. *British Medical Journal* **325**:177-178.

Robinson, L.D. and Jewell, N.P. (1991) Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **58**:227-240.

Robinson, W.S. (1950) Ecological correlations and the behavior of individuals. *American Socialogical Review* **15**:351-357.

Rogowski, J.A., Horbar, J.D., Staiger, D.S., Kenny, M., Carpenter, J. and Geppert, J. (2004) Indirect vs direct hospital quality indicators for very low-birth-weight infants. *Journal of the*

American Medical Association **291**(2):202-209.

Roman, E., Doyle, P.B.V., Alberman, E. and Pharoah, P. (1978) Fetal loss, gravidity, and pregnancy order. *Early Human Development* **2**:131-138.

Romero, R., Chaiworapongsa, T. and Espinoza, J. (2003) Micronutrients and intrauterine infection, preterm birth and the fetal inflammatory response syndrome. *Journal of Nutrition* **133**:1668S-1673S.

Roques, F., Michel. P., Goldstone, A.R. and Nashef, S.A.M. (2003) The logistic EuroSCORE. *European Heart Journal* **24**:1-2.

Rosen, M.G. and Dickinson, J.C. (1993) The paradox of electronic fetal monitoring: more data may not enable us to predict or prevent infant neurologic morbidity. *American Journal of Obstetrics and Gynecology* **168**:745-751.

Rosenberg, K.D., Desai, R.A., Na, Y., Kan, J.L. and Schwartz, L. (2001) The effect of surfactant on birthweight-specific neonatal mortality rate, New York City. *Annals of Epidemiology* **11**:337-341.

Rosenthal, G.E. and Harper, D.L. (1994) Cleveland health quality choice: a model for collaborative community-based outcome assessment. *Joint Commission Journal on Quality Improvement* **20**:425-422.

Rothschild, V. and Logothetis, N. (1986) Probability Distributions. New York, Wiley.

Rowan, K.M., Kerr, J.H., Major, E., McPherson, K., Short, A. and Vessey, M.P. (1993a) Intensive Care Society's APACHE II study in Britain and Ireland--II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *British Medical Journal* **307** (6910):977-81.

Rowan, K.M., Kerr, J.H., Major, E., McPherson, K., Short, A. and Vessey, M.P. (1993b) Intensive Care Society's APACHE II study in Britain and Ireland--I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. *British Medical Journal* **307** (6910):972-7.

Royal College of Obstetricians and Gynaecologists (2001) The use of electronic fetal monitoring: the use and interpretation of cardiotocography in intrapartum fetal surveillance.

Evidence based clinical guideline number 8. London, Royal College of Obstetricians and Gynaecologists.

Royal College of Obstetricians and Gynaecologists (2004) Guideline No. 7: Antenatal corticosteroids to prevent respiratory distress syndrome. London, RCOG.

Royal College of Paediatrics and Child Health (1997) Withholding or withdrawing life saving treatment in children. A framework for practice. London, Royal College of Paediatrics and Child Health.

Royal Statistical Society Working Party on Performance Monitoring in the Public Services (2004) Performance indicators: good, bad, and ugly. London, Royal Statistical Society.

Royston, P., Ambler, G. and Sauerbrei, W. (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* **28**:964-974.

Rubin, D.B. (1997) Estimating effects from large data sets using propensity scores. *Annals of Internal Medicine* 127(8):757-763.

Ruttimann, U.E. and Pollack, M.M. (1996) Variability in Duration of Stay in Pediatric Intensive Care Units: a Multiinstitutional Study. *Journal of Pediatrics* **128**:35-44.

Ryan, M. and Farrar, S. (2000) Using Conjoint Analysis to Elicit Preferences for Health Care. *British Medical Journal* **320**:1530-1533.

Sackett, D.L., Deeks, J.J. and Altman, D.G. (1996) Down with odds ratios! *Evidence Based Medicine* **1**:164-165.

Saling, E. (1996) Comments on past and present situation of intensive monitoring of the fetus during labor. *Journal of Perinatal Medicine* **24**:7-13.

Sankaran, K., Chien, L.-Y., Walker, R., Seshia, M., Ohlsson, A., Lee, S.K. and the Canadian Neonatal Network (2002) Variation in mortality rates among Canadian neonatal intensive care units. *Canadian Medical Association Journal* **166**:173-178.

SAS Institute Inc. (1999) SAS/STAT® User's Guide, Version 8. Cary, NC, SAS Institute Inc.

Savitz, D.A., Terry, J.W., Dole, N., Thorp, J.M., Siega-Riz, A.M. and Herring, A.H. (2002)

Comparison of Pregnancy Dating by Last Menstrual Period, Ultrasound Scanning, and Their Combination. *American Journal of Obstetrics and Gynecology* **187**:1660-1666.

Scally, G. and Donaldson, L.J. (1998) Clinical Governance and the Drive for Quality Improvement in the New NHS in England. *British Medical Journal* **317**:61-65.

Schneider, E.C. and Epstein, A.M. (1996) Influence of cardiac-surgery performance reports on referral practices and access to care. A survey of cardiovascular specialists. *New England Journal of Medicine* **335** (4):251-256.

Scottish Neonatal Consultants' Collaborative Study Group and the International Neonatal Network (1995) CRIB (Clinical Risk Index for Babies), mortality, and impairment after neonatal intensive care. Scottish Neonatal Consultants' Collaborative Study Group and the International Neonatal Network. *Lancet* **345** (8956):1020-1022.

Seagroatt, V. and Goldacre, M.J. (2004) Hospital mortality league tables: influence of place of death. *British Medical Journal* **328**:1235-1236.

Seeds, J.W. and Peng, T. (1998) Impaired Growth and Risk of Fetal Death: Is the Tenth Percentile the Appropriate Standard? *American Journal of Obstetrics and Gynecology* **178**:658-669.

Selbmann, H.K., Warncke, W. and Eissner, H.J. (1982) Comparison of hospitals supporting quality assurance. *Methods of Information in Medicine* **21**(2):75-80.

Selvin, H.C. (1958) Durkheim's suicide and problems of empirical research. *American Journal of Sociology* **63**:607-619.

SF-36 (2005)
Available at http://www.sf-36.org/

Shankaran, S., Fanaroff, A.A., Wright, L.L., Stevenson, D.K., Donovan, E.F., Ehrenkranz, R.A., Langer, J.C., Korones, S.B., Stoll, B.J., Tyson, J.E., Bauer, C.R., Lemons, J.A., Oh, W. and Papile, L.A. (2002) Risk factors for early death among extremely low-birth-weight infants. *American Journal of Obstetrics and Gynecology* **186**:796-802.

Shann, F., Pearson, G., Slater, A. and Wilkinson, K. (1997) Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care. *Neonatal and Pediatric*

Intensive Care **23**:201-207.

Shaw, C.D. (1997) Health-care league tables in the United Kingdom. *Journal of Quality in Clinical Practice* **17**(4):215-9.

Sheay, W., Ananth, C.V. and Kinzler, W.L. (2004) Perinatal mortality in first- and second-born twins in the United States. *Obstetrics and Gynecology* **103**:63-70.

Sheldon, T. (1998) Promoting health care quality: what role performance indicators? *Quality in Health Care* **7**:S45-S50.

Sheldon, T. (2001) Dutch doctors change policy on treating preterm babies. *British Medical Journal* **322**:1383.

Sheldon, T. (2004) Dutch doctors call for new approach to reporting "mercy killings". *British Medical Journal* **329**:591.

Sherlaw-Johnson, C., Lovegrove, J., Treasure, T. and Gallivan, S. (2000) Likely variations in perioperative mortality associated with cardiac surgery: when does high mortality reflect bad practice? *Heart* **84**:79-82.

Shinwell, E.S., Blickstein, I., Lusky, A. and Reichman, B. (2004) Effect of birth order on neonatal morbidity and mortality mong very low birthweight twins: a population based study. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F145-F148

Shiu, M.F. (2002) Performance league tables - use of language should be more careful in describing league tables. *British Medical Journal* **324**:542

Shwartz, M., Iezzoni, L.I., Ash, A.S. and Mackiernan, Y.D. (1996) Do severity measures explain differences in length of hospital stay? The case of hip fracture. *Health Services Research* **31**(4):365-85.

Signorini, D.F. and Weir, N.U. (1999) Any variability in outcome comparisons adjusted for case mix must be accounted for [letter]. *British Medical Journal* **318**(7176):128

Silber, J.H., Rosenbaum, P.R., Schwartz, J.S., Ross, R.N. and Williams, S.V. (1995) Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery. *Journal of the American Medical Association* **274**(4):317-323.

Silcock, H. (1959) The comparison of occupational mortality rates. *Population Studies* **13**:183-192.

Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B - Methodological* **13**:238-241.

Simpson, J.M., Evans, N., Gibberd, R.W., Heuchan, A.M., Henderson-Smart, D.J., on behalf of the Australia and New Zealand Neonatal Network (2003) Analysing differences in clinical outcomes between hospitals. *Quality and Safety in Health Care* **12**:257-262.

Sinclair, J.C. and Bracken, M.B. (1994) Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* **47**:881-889.

Singh, J.K., Khoshnood, B., Hsieh, H.-L., Sriram, S. and Lee, K. (1997) Are differences in birth weight-specific neonatal mortality between racial/ethnis groups explainable by differences in Apgar scores? *Pediatric Research* **41**:210

Sinkin, R.A., Cox, C. and Phelps, D.L. (1990) Predicting Risk for Bronchopulmonary Dysplasia - Selection Criteria for Clinical-Trials. *Pediatrics* **86**:728-736.

Skjaerven, R., Gjessing, H.K. and Bakketeig, L.S. (2000) Birthweight by Gestational Age in Norway. *Acta Obstetricia Et Gynecologica Scandinavica* **79**:440-449.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B – Methodological* **55**(1):3-23.

Snijders, T.A.B. and Bosker, R.J. (1999) Multilevel Analysis: An introduction to basic and advanced multilevel modeling. London, Sage.

So, Y. (1993) A tutorial on logistic regression. SAS Institute Inc., Proceedings of the Eighteenth Annual SAS Users Group International Conference. Cary NC, SAS Institute Inc.:1290

Society of Cardiothoracic Surgeons of Great Britain and Ireland (2002) National Audit Cardiac Surgical Database Report 2000-2001. Henley-on-Thames, Dendrite Clinical Systems Ltd.

Society of Cardiothoracic Surgeons of Great Britain and Ireland. (2004) Available at http://ctsnet.org/doc/6144

Society of Cardiothoracic Surgeons of Great Britain and Ireland. (2005) http://www.scts.org/index.cfm?ukcardiacreg=yes

Sola, A. and Chow, L.C. (1999) The coming of (gestational) age for preterm infants. *Journal of Pediatrics* **135**:137-139.

Spiegelhalter, D. (2002) Funnel plots for institutional comparison. *Quality and Safety in Health Care* **11**:390-391.

Spiegelhalter, D. (2003) Ranking institutions. *Journal of Thoracic & Cardiovascular Surgery* **125**:1171-1173.

Spiegelhalter, D. (2005) Funnel plots for comparing institution performance. *Statistics in Medicine* **24**:1185-1202.

Spiegelhalter, D.J., Abrams, K.R. and Myles, J.P. (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, Wiley.

Spiegelhalter, D.J. (1999) Surgical audit: statistical lessons from Nightingale and Codman. *Journal Of The Royal Statistical Society Series A - Statistics In Society* **162**(Pt1):45-58.

Spiegelhalter, D.J., Aylin, P., Best, N.G., Evans, S.J.W. and Murray, G.D. (2002) Commissioned analysis of surgical performance by using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society A - Statistics in Society* **165**:191-231.

Spiegelhalter, D.J., Freedman, L.S. and Parmar, M.K.B. (1994) Bayesian approaches to randomised trials. *Journal Of The Royal Statistical Society Series A - Statistics In Society* **157**:357-387.

Spiegelhalter, D.J., Grigg, O., Kinsman, R. and Treasure, T. (2003) Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *International Journal for Quality in Health Care* **15**(1):7-13

Spiegelhalter, D.J., Myles, J.P., Jones, D.R. and Abrams, K.R. (1999a) Methods in health service research: an introduction to Bayesian methods in health technology assessment.

British Medical Journal **319**:508-512.

Spiegelhalter, D.J., Thomas, A. and Best, N.G. (1999b) WinBUGS Version 1.2 User Manual. Cambridge, MRC Biostatistics Unit.

St George's Hospital (2005)
Available at http://www.st-georges.org.uk/cardiacinfo.asp

Stark, J., Gallivan, S., Lovegrove, J., Hamilton, J.R.L., Monro, J.L., Pollock, J.C.S. and Watterson, K.G. (2000) Mortality rates after surgery for congenital heart defects in children and surgeons' performance. *Lancet* **355**:1004-1007.

Steel, R.G.D. and Torrie, J.H. (1960) Principles and Procedures of Statistics. New York, McGraw-Hill.

Sterne, J.A.C., Egger, M. and Davey Smith, G. (2001) Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal* **323**:101-105.

Stevens, S.M., Richardson, D.K., Gray, J.E., Goldmann, D.A. and McCormick, M.C. (1994) Estimating neonatal mortality risk: an analysis of clinicians' judgments. *Pediatrics* **93**(6 Pt 1):945-950.

Stevenson, D.K., Verter, J., Fanaroff, A.A., Oh, W., Ehrenkranz, R.A., Shankaran, S., Donovan, E.F., Wright, L.L., Lemons, J.A., Tyson, J.E., Korones, S.B., Bauer, C.R. and Stoll, B.J. (2000) Sex differences in outcomes of very low birthweight infants: the newborn male disadvantage. *Archives of Disease in Childhood Fetal & Neonatal Edition* **83**:F182-F185

Stuart, A. and Ord, J.K. (1994) Kendall's Advanced Theory of Statistics. Vol. 1: Distribution Theory. 6th edn, London, Hodder Arnold.

Sullivan, S.A. and Newman, R. (2004) Prediction and prevention of preterm deliveries in multiple gestations. *Clinical Obstetrics and Gynecology* **47**:203-215.

Suresh, G.K. and Soll, R.F. (2001) Current surfactant use in premature infants. *Clinics in Perinatology* **28**:671-694.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. and Song, F. (2000) Methods for meta-analysis in medical research. Chichester, Wiley.

Takemura, Y., Kikuchi, S. and Inaba, Y. (1998) Epidemiologic study of the relationship between schistosomiasis due to schistosoma japonicum and liver cancer/cirrhosis. *American Journal of Tropical Medicine and Hygiene* **59**:551-556.

Tarnow-Mordi, W. (1997) Commentary to Maier RF, M Rey, BC Metze, M Obladen: Comparison of mortality risk: a score for very low birthweight infants. *Archives of Archives of Disease in Childhood Fetal & Neonatal Edition* **76**:F150-F151.

Temmerman, M., Verstraelen, H., Marten, G. and Bekaert, A. (2004) Delayed childbearing and maternal mortality. *European Journal of Obstetrics, Gynecology and Reproductive Biology* **114**:19-22.

Templeton, S. and Rogers, L. (Jun 20:2004) Babies that live after abortions are left to die. *The Sunday Times*

The EuroQol Group (1990) EuroQol: a new facility for the measurement of health-related quality of life. *Health Policy* **16**:199-208.

The Guardian (Oct 31:2003)

The International Neonatal Network (1993) The CRIB (Clinical Risk Index for Babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. *Lancet* **342**(8865):193-8.

The New Oxford English Dictionary (1998) Oxford, Oxford University Press.

The Shipman Enquiry (2002) First Report. Volume One: Death Disguised. London, HMSO.

The Sunday Times (May 16:2004a).

The Sunday Times (May 16:2004b) The Sunday Times Good Hospital Guide.

The Times (Nov 19:2001) Hospital consultants' guide. *The Times* Suppl. part I edn:22-23. Suppl. part I.

The Times (Nov 3:2003)

The Trent Infant Mortality and Morbidity Studies (2000) Leicester, The Trent Infant Mortality and Morbidity Studies.

The Trent Infant Mortality and Morbidity Studies (2003) Leicester, The Trent Infant Mortality and Morbidity Studies.

Thomas, J.W., Holloway, J.J. and Guire, K.E. (1993) Validating risk-adjusted mortality as an indicator for quality of care. *Inquiry* **30**(1):6-22.

Tibby, S.M., Taylor, D., Festa, M., Hanna, S., Hatherill, M., Jones, G., Habibi, P., Durward, A. and Murdoch, I.A. (2002) A comparison of three scoring systems for mortality risk among retrieved intensive care patients. *Archives of Disease in Childhood Fetal & Neonatal Edition* **87**:F421-F425.

Tietz, N.W. (1986) Textbook of clinical chemistry. Philadelphia, W. B. Saunders Compary .

Townsend, P., Phillimore, P. and Beattie, A. (1988) Health and deprivation: inequality in the North. London, Routledge.

Trudinger, B.J. (1999) Doppler ultrasonography and fetal well-being. In: Reece, E.A. and Hobbins, J.C., (Eds.) *Medicine of the fetus & mother*, 2nd edn. Philadelphia: Lippincott-Raven, pp. 753-777.

Trudinger, B.J., Cook, C.M., Giles, W.B., Ng, S., Fong, E., Connelly, A. and Wilcox, W. (1991) Fetal umbilical artery velocity waveforms and subsequent neonatal outcome. *British Journal of Obstetrics and Gynaecology* **98**:378-384.

Tucker, J., Parry, G., Fowlie, P.W., McGuire, W. (2004) Organisation and delivery of perinatal services. *British Medical Journal* **329**(7468):730-732.

Tukey, J.W. (1969) Analyzing data: sanctification or detective work? *American Psychologist* **24**:83-91.

Tunon, K., Eik-Nes, S.H. and Grottum, P. (1999) Fetal outcome in pregnancies defined as post-term according to the last menstrual period estimate, but not according to the ultrasound estimate. *Ultrasound in Obstetrics & Gynecology* **14**:12-16.

Turner, R.M., Omar, R.Z. and Thompson, S.G. (2001) Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* **20**:453-472.

Tybulewicz, A.T., Clegg, S.K., Fonfé, G.J. and Stenson, B.J. (2004) Preterm meconium

staining of the amniotic fluid: associated findings and risk of adverse clinical outcome. *Archives of Disease in Childhood Fetal & Neonatal Edition* **89**:F328-F330

UK Neonatal Staffing Study Group (2002) Patient Volume, Staffing, and Workload in Relation to Risk- Adjusted Outcomes in a Random Stratified Sample of UK Neonatal Intensive Care Units: a Prospective Evaluation. *Lancet* **359**:99-107.

Van Den Berg, B.J. and Yerushalmy, J. (1966) The relationship of the rate of intrauterine growth of infants of low birth weight to mortality, morbidity, and congenital anomalies. *Journal of Pediatrics* **69**:531-545.

Vass, A. (2001) Doctors urge caution in interpretation of league tables. *British Medical Journal* **323**:1205.

Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**(433):217-221.

Verloove-Vanhorick, S.P., Verwey, R.A., Brand, R., Gravenhorst, J.B., Keirse, M.J.N.C. and Ruys, J.H. (1986) Neonatal-mortality risk in relation to gestational-age and birth-weight - results of a national survey of preterm and very-low-birth-weight infants in the Netherlands. *Lancet* **1**:55-57.

Vintzileos, A.M., Ananth, C.V., Smulian, J.C., Scorza, W.E. and Knuppel, R.A. (2002) The impact of prenatal care on neonatal deaths in the presence and absence of antenatal high-risk conditions. *American Journal of Obstetrics and Gynecology* **186**:1011-1016.

Vyas, J., Field, D., Draper, E.S., Woodruff, G., Fielder, A.R., Thompson, J., Shaw, N.J., Clark, D., Gregson, R., Burke, J. and Durbin, G. (2000) Severe retinopathy of prematurity and its association with different rates of survival in infants of less than 1251 g birth weight. *Archives of Disease in Childhood Fetal & Neonatal Edition* **82**:F145-F149

Wacholder, S. (1986) Binomial regression in GLIM: estimating risk ratios and risk differences. *American Journal of Epidemiology* **123**:174-180.

Waller, D.K., Spears, W.D., Gu, Y. and Cunningham, G.C. (2000) Assessing number-specific error in the recall of onset of last menstrual period. *Paediatric and Perinatal Epidemiology* **14**:263-267.

Wariyar, U., Tin, W. and Hey, E. (1997) Gestational assessment assessed. *Archives of Disease in Childhood Fetal & Neonatal Edition* **77**:F216-F220

Waugh, J., Johnson, A. and Farkas, A. (2001) Analysis of cord blood gas at delivery: questionnaire study of practice in the United Kingdom. *British Medical Journal* **323**:727

Weinberger, B., Anwar, M., Hegyi, T., Hiatt, M., Koons, A. and Paneth, N. (2000) Antecedents and neonatal consequences of low apgar scores in preterm newborns - a population study. *Archives of Pediatrics & Adolescent Medicine* **154**:294-300.

Wen, S.W., Kramer, M.S., Liu, S., Dzakpasu, S. and Sauve, R. (2000) Infant mortality by gestational age and birth weight in Canadian Provinces and Territories:1990-1994 births. *Chronic Diseases in Canada* **21**:14-22.

Whitelaw, A. (1986) Death as an option in neonatal intensive-care. *Lancet* **2**:328-331.

Wilcox, A. and Russell, I. (1990) Why small black infants have a lower mortality rate than small white infants: the case for population-specific standards for birth weight. *Journal of Pediatrics* **116**:7-10.

Wilcox, A.J. (2001) On the importance - and the unimportance - of birthweight. *International Journal of Epidemiology* **30**:1233-1241.

Wilcox, A.J. and Russell, I.T. (1983) Perinatal mortality: standardizing for birthweight is biased. *American Journal of Epidemiology* **118**:857-864.

Wilcox, A.J. and Russell, I.T. (1986) Birthweight and perinatal mortality: III. Towards a new method of analysis. *International Journal of Epidemiology* **15**:188-196.

Wilson, P., Smoley, S.R. and Werdegar, D. (1996) Second report of the California Hospital Outcomes Project: acute myocardial infarction, Volume One: Study overview and results summary. Sacramento, Calif., Office of Statewide Health Planning and Development.

Winkler, C.L., Hauth, J.C., Tucker, J.M., Owen, J. and Brumfield, C.G. (1991) Neonatal complications at term as related to the degree of umbilical artery acidemia. *American Journal of Obstetrics and Gynecology* **164**:637-641.

Winston, R. (1998) League tables of in vitro fertilisation clinics misinform patients [letter].

British Medical Journal **317**(7172):1593-1594.

Wolf, H., Schaap, A.H., Bruinse, H.W., Smolders-de Haas, H., van Ertbruggen, I. and Treffers, P.E. (1999) Vaginal delivery compared with caesarean section in early preterm breech delivery: a comparison of long term outcome. *British Journal of Obstetrics and Gynaecology* **106**:486-491.

Wolfe, C.D.A., Tilling, K., Beech, R. and Rudd, A.G. (1999) Variations in case fatality and dependency from stroke in Western and Central Europe. *Stroke* **30**:350-356.

World of quotes (2005)
Available at http://www.worldofquotes.com/author/Albert-Einstein/1/

Yang, H., Kramer, M.S., Platt, R.W., Blondel, B., Breart, G., Morin, I., Wilkins, R. and Usher, R. (2002) How does early ultrasound scan estimation of gestational age lead to higher rates of preterm birth? *American Journal of Obstetrics and Gynecology* **186**:433-437.

Yerushalmy, J. (1970) Relation of birth weight, gestational age, and the rate of intrauterine growth to perinatal mortality. *Clinical Obstetrics and Gynecology* **13**:107-129.

Young, A.C. (1993) Consultants' league tables. *British Medical Journal* **307**(6911):1070.

Younge, P.A., Coats, T.J., Gurney, D. and Kirk, C.J.C. (1997) Interpretation of the W Statistic: application to an integrated trauma tystem. *Journal of Trauma: Injury Infection and Critical Care* **43**:511-515.

Yudkin, P. (1980) Pregnancy order and reproductive loss. *British Medical Journal* **280**:715-716.

Yudkin, P.L., Aboualfa, M., Eyre, J.A., Redman, C.W.G. and Wilkinson, A.R. (1987) Influence of elective preterm delivery on birthweight and head circumference standards. *Archives of Disease in Childhood* **62**:24-29.

Yule, G.U. (1934) On some points relating to vital statistics, more especially statistics of occupational mortality. *Journal of the Royal Statistical Society* **97**:1-84.

Zardo, M.S. and Procianoy, R.S. (2003) Comparison between different mortality risk scores in a neonatal intensive care unit. *Revista de Saúde Pública* **37**:591-596.

Zhang, J. and Bowes, W.A. (1995) Birth-weight-for-gestational-age patterns by race, sex, and parity in the United-States population. *Obstetrics and Gynecology* **86**:200-208.

Zhang, P. (2003) Multiple imputation: theory and method. *International Statistical Review* **71**:581-592.

Zhou, H. and Romano, P.S. (1997) Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine* **16**:1301-1303.

Zullini, M.T., Bonati, M. and Sanvito, E. (1997) Survival at nine neonatal intensive care units in Sao Paulo, Brazil. Paulista Collaborative Group on Neonatal Care. *Pan American Journal of Public Health* **2** (5):303-309.