Title: A flexible parametric approach to examining spatial variation in relative survival

Authors: Susanna M Cramb[a,b], Kerrie L Mengersen[b,c], Paul C Lambert[d], Louise M Ryan[e], Peter D Baade[a,f,g]

[a] Cancer Council Queensland, Brisbane, Australia.

[b] ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Australia

[c] Cooperative Research Centre for Spatial Information, Australia.

[d] Department of Health Sciences, University of Leicester, UK.

[e] ARC Centre of Excellence for Mathematical and Statistical Frontiers, University of Technology, Sydney, Australia.

[f] School of Public Health and Social Work, Queensland University of Technology (QUT), Brisbane, Australia.

[g] Griffith Health Institute, Griffith University, Gold Coast, Australia.

Corresponding author: Dr Susanna Cramb, Cancer Council Queensland, PO Box 201, Spring Hill, QLD 4004, Australia; E-mail: susannacramb@cancerqld.org.au; Telephone: +61-7-36345350; Facsimile: +61-7-32598527.

Sources of support: PDB was supported by an Australian National Health and Medical Research Council Career Development Fellowship (#1005334). KLM acknowledges support from the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. LMR and KLM acknowledge support from the ARC Centre of Excellence in Mathematical and Statistical Frontiers. The views expressed in this paper are those of the authors and not of any funding body.

Conflict of Interest Statement

Nil. All funding sources have been acknowledged.

Keywords: flexible parametric; relative survival; Bayesian; cancer; Australia; small-area

A flexible parametric approach to examining spatial variation in relative survival

Abstract

Most of the few published models used to obtain small-area estimates of relative survival are based on a generalized linear model with piecewise constant hazards under a Bayesian formulation. Limitations of these models include the need to artificially split the time scale, restricted ability to include continuous covariates, and limited predictive capacity. Here, an alternative Bayesian approach is proposed: a spatial flexible parametric relative survival model. This overcomes previous limitations by combining the benefits of flexible parametric models: the smooth, well-fitting baseline hazard functions and predictive ability, with the Bayesian benefits of robust and reliable small-area estimates. Both spatially structured and unstructured frailty components are included. Spatial smoothing is conducted using the intrinsic conditional autoregressive prior. The model was applied to breast, colorectal and lung cancer data from the Queensland Cancer Registry across 478 geographical areas. Advantages of this approach include the ease of including more realistic complexity, the feasibility of using individual-level input data, and the capacity to conduct overall, cause-specific and relative survival analysis within the same framework. Spatial flexible parametric survival models have great potential for exploring small-area survival inequalities, and we hope to stimulate further use of these models within wider contexts.

1. Introduction

Spatial analyses of routinely collected cancer data are being increasingly used to provide insight to disease etiology and to inform decisions regarding health care disparities [1]. These analyses typically report on variation in cancer incidence and mortality [2]. While providing important information on diagnostic and end of life care requirements, these endpoints provide limited information on the effectiveness of cancer-related health care systems. Spatial survival analyses provide greater opportunity to assess the geographical variation in the

effectiveness of health services as they reflect both diagnostic and patient management components [3].

Cancer survival may be reported as overall survival, where deaths from any cause are included, or an estimate of net survival. Net survival is the survival that would be seen if the cancer under study was the only possible cause of death. Net survival is estimated using either cause-specific survival, where the recorded cause of death determines deaths due to cancer, or relative survival, where deaths from any cause among patients are compared against background population mortality rates. When using population-based data such as from a cancer registry, relative survival is often the preferred method for measuring net survival, as the accuracy of the recorded cause of death may be uncertain [4].

Few small-area analyses have used relative survival. Fairley *et al* [5] examined prostate cancer in a region of the UK, Cramb *et al* [6] examined a range of cancers across Queensland, Australia, and Saez *et al* [7] investigated breast cancer in a region of Spain. Each of these analyses used a generalized linear model (GLM) with a modified link function, piecewise constant hazards and spatial frailties within a Bayesian framework. Reliable small area estimates were obtained by using prior distributions which smoothed estimates across adjacent areas. While the GLM has been recommended for modelling relative survival [4], the disjointed piecewise hazards are biologically implausible. Hennerfeind *et al* sought to overcome this by using a similar model which incorporated splines to smooth the piecewise constant hazards [8]. However, while the use of splines has the potential to provide a better fit to the hazards function, the resulting calculations for this model are computationally-intensive [8].

Fully parametric models have several advantages over piecewise linear approaches. For instance, the time scale does not need to be artificially split, it is more feasible to model individual-level data rather than aggregating over covariates of interest, and it is simpler to obtain smooth survival or hazard function predictions [9]. However, the standard parametric formulations, such as the Weibull, log-logistic or log-normal distributions, assume a linear relationship between a specific transformation of the survival function and log survival time [9], which often results in poorly fitting models.

Flexible parametric models incorporate the advantages of standard parametric models with nonlinear functions for modelling the baseline hazard, enabling improved model fit. One

version of these flexible parametric models is the model proposed by Royston and Parmar which uses restricted cubic splines to model the log cumulative baseline hazard [9]. Nelson *et al* [10] extended this flexible parametric model to the relative survival context. However, these models have not been previously used for small-area survival analyses.

The purpose of this paper is to introduce an alternative method for geographic analysis of cancer survival data: the spatial flexible parametric relative survival model. We extend Nelson's model [10] to the spatial context by incorporating random effects that allow for spatial correlation between areas. This was implemented using a Bayesian framework. We apply this new model to three common cancers in Australia: breast cancer (high survival), colorectal cancer (moderate survival), and lung cancer (low survival). Our focus is on the practical implementation, predictive capacity and interpretability of results. In sections 2 and 3, details of the proposed model are presented, along with the data and analyses. Model assessment is described in section 4, and results presented in section 5, focusing on the predictive options available under the flexible parametric formulation. Finally, section 6 discusses the implications of these new models.

2. Model Formulation

Relative survival partitions the total mortality rate (overall hazard, $h(t)$) into that resulting from the disease of interest (excess hazard, $\lambda(t)$) and that due to other causes (expected hazard ($h^*(t)$, estimated from population mortality rates) [9]. This is also known as an additive hazards model since it can be expressed as:

$$h(t) = h^*(t) + \lambda(t) \quad (1)$$

The relative survival function for an individual with covariate vector x can be represented as:

$$\ln(-\ln R(t;x)) = \ln(\Lambda(t)) = \ln(\Lambda_0(t)) + x\beta, \quad (2)$$

where $R(t;x)$ is the relative survival function, $\Lambda(t)$ is the cumulative excess hazard, which is the integrated form of $\lambda(t)$ in (1), $\Lambda_0(t)$ is the cumulative baseline excess hazard (the cumulative excess hazard when all covariates are 0) and $\beta = \beta_1, \ldots, \beta_K$ and represents the vector of coefficients relating to covariates x.

The log cumulative baseline excess hazard ($\ln(\Lambda_0(t))$ is modelled via restricted cubic splines

[11] as a function of log time. When at least one interior knot is specified, the spline includes a constant term, γ_0, a linear function of log time with parameter γ_1, and for each interior knot m=1,…,M, a basis function, $z_m(t)$, with parameter γ_{m+1}, as follows:

$$\ln\big(\Lambda_0(t)\big) = \gamma_0 + \gamma_1 \ln(t) + \gamma_2 z_1(t) + \cdots + \gamma_{M+1} z_M(t). \quad (3)$$

The cubic basis functions $z_1(t), \ldots, z_M(t)$ are calculated as:

$$z_m(t) = (\ln(t) - k_m)_+^3 - \frac{k_{\max} - k_m}{k_{\max} - k_{\min}} (\ln(t) - k_{\min})_+^3 - \left(1 - \frac{k_{\max} - k_m}{k_{\max} - k_{\min}}\right) (\ln(t) - k_{\max})_+^3 \quad (4)$$

with M interior knots $k_1, \ldots, k_M$ and two boundary knots ($k_{\min}$ and $k_{\max}$), as per Royston and Lambert [9]. The + subscript indicates that negative values are truncated at zero. Note that if no interior knots are specified, (3) will revert to a standard Weibull model with $\ln(\Lambda_0(t)) = \gamma_0 + \gamma_1 \ln(t)$.

When this model is implemented in Stata (via stpm2 [12]) or R (via package 'flexsurv' [13]), the number and location of knots must be pre-selected along with data relating to the population-based expected hazard. The number of interior knots may be selected using measures of fit such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or graphical plots of model fit. The two boundary knots $k_{\min}$, $k_{\max}$ are placed at the smallest and largest uncensored log survival-times, respectively. The default for these models is to position interior knots using empirical centiles of the distribution of log event times, which allows data to be more closely modelled in regions of greater data density [9]. For example, 1 interior knot is positioned at the median, 2 interior knots are positioned at the 33rd and 67th centiles, 3 interior knots are positioned at the 25th, 50th and 75th centile, and so on. Maximum likelihood is used to estimate the spline parameters and the log hazard ratios. Log-likelihood functions are maximised using the Newton Raphson technique [10].

We introduce the spatial flexible parametric relative survival model by extending Nelson's model in (2) with the cumulative baseline excess hazard specified as in (3) to include additional spatial frailty terms. Using notation similar to that used by Gelman and Hill [14], suppose the ith individual with covariate x_i lives in area j (represented as $j[i]$), then the log cumulative hazard can be written as follows:

$$\ln\Big(\Lambda\big(t; \mathbf{x}_i; u_{j[i]}; v_{j[i]}\big)\Big) = \gamma_0 + \gamma_1 + \gamma_2 z_1(t) + \cdots + \gamma_{M+1} z_M(t) + \mathrm{x}_i\beta + u_{j[i]} + v_{j[i]} \quad (5)$$

where $u_{j[i]}$ and $v_{j[i]}$ are random effects representing the spatial and uncorrelated heterogeneity, respectively, in *j*=1,…J areas. The $v_{j[i]}$ terms receive independent normal distributions and the $u_{j[i]}$ are assumed to follow an intrinsic conditional autoregressive distribution [15].

A Bayesian framework was used to enable the smoothing of estimates over regions. Since the additional complexity of priors and hyperpriors precludes an analytical solution, Markov Chain Monte Carlo (MCMC) sampling was used to obtain estimates.

Probability distributions were placed on each parameter, with Gaussian distributions expressed as N(mean, variance) as follows:

$\gamma_p \sim \mathrm{N}(0, 10^6)$ where *p*=0,...,*P* and *P*=*M* + 1 where *M* is the maximum number of interior knots

$\beta \sim \mathrm{N}(0, 10^6)$

$u_j|\mathbf{u_{-j}} = \mathrm{N}\left(\bar{\mu}_j, \frac{\sigma_u^2}{n_j}\right)$ where $\bar{\mu}_j$ is the average of the neighboring regions of area *j*

$\mathbf{u_{-j}} = (u_1, \ldots u_{j-1}, u_{j+1}, \ldots, u_J)$

$v_j \sim \mathrm{N}(0, \sigma_v^2)$

$\sigma_u \sim \mathrm{Uniform}(0.01,20)$

$\sigma_v \sim \mathrm{Uniform}(0.01,20)$

Hyperpriors were provided as a uniform distribution on the standard deviation [16]. A popular choice to encourage the influence of the data is to use vague prior distributions, and here most model parameters were given vague normal distributions, except for *u*, as the spatially structured term. The intrinsic CAR distribution [15] prior specified on *u* locally smooths the data across neighbors (n_j=number of neighbors of region *j*), defined as areas with first-order contiguity (adjacent boundaries). As Queensland has many small islands (which have no adjacent boundaries), these areas had their default neighborhood structure adjusted to incorporate nearby areas.

Although (5) results in proportional hazards, alternative formulations are possible, including proportional odds models. The equivalent to (5) under proportional odds is:

$$\text{logit}\left(1 - R\left(t; \mathbf{x}_i; u_{j[i]}; v_{j[i]}\right)\right) = \text{logit}\left(1 - R_0(t)\right) + \text{x}_i\beta + u_{j[i]} + v_{j[i]} \quad (6)$$

where $R\left(t; \mathbf{x}_i; u_{j[i]}; v_{j[i]}\right)$ represents relative survival for individual *i* in area *j* given covariates x_i, $R_0(t)$ is the baseline relative survival function, the relative survival when all covariates equal zero and other terms are as before. Note that $\text{logit}\left(1 - R_0(t)\right)$ is modelled as a restricted cubic spline just as for the log cumulative baseline hazard in (3).

In a slight abuse of notation but to simplify exposition we write $R\left(t; \mathbf{x}_i; u_{j[i]}; v_{j[i]}\right)$ as $R(t_i)$. The resulting contribution to the log-likelihood for the *i*th individual when allowing for late entry at t_{0i} can then be written as:

$$\ln L_i = \delta_i \ln\{h^*(t_i) + \lambda(t_i)\} + \ln S^*(t_i) + \ln R\,(t_i) - \ln S^*(t_{0i}) - \ln R\,(t_{0i}) \quad (7)$$

where δ_i is a death indicator with 0 representing censored and 1 representing death, $h^*(t_i)$ and $\lambda(t_i)$ are as specified in (1), $S^*(t_i)$ is the expected survival at the time of death or censoring and $S^*(t_{0i})$ is the expected survival during the time period prior to the time period of interest (as no model parameters are included in these terms, they can be excluded when maximising the likelihood), $R(t_i)$ is the relative survival from (6), and $R(t_{0i})$ is the relative survival before the time period of interest. As the models use the cumulative form, the excess hazard $\lambda(t_i)$ is obtained as the first derivative of $\ln(\Lambda(t))$. Further details are available in Appendix A.

The choice of proportional hazards or proportional odds may be determined using the same measures as for choosing the number of knots (AIC, BIC or graphical plots of model fit), often in conjunction with the interpretability of estimates and audience needs. Hazard ratios may be easier to interpret due to their widespread use, but the proportional odds model assumes hazard ratios decrease as time from diagnosis increases, which is often sensible for cancer prognostic effects.

3. Data and Analysis

Breast (ICD-O3 C50), colorectal (ICD-O3 C18-C20,C218) and lung (ICD-O3 C33-C34) cancer data for patients aged <90 years diagnosed from 1997 to 2011 were obtained from the Queensland Cancer Registry (QCR), a population-based registry that covers the entire state of Queensland [17]. The QCR conducts routine data linkage with the Australian National Death Index to determine the survival status of all cancer patients. Ethical approval was obtained from the Queensland Health Central Office Human Research Ethics Committee (HREC/09/QHC/25).

The following variables collected by the QCR have been recognized as important prognostic indicators for both breast [18] and colorectal [19] cancers: patient age, sex, tumor stage at diagnosis (these are all included within the covariate matrix), and geographical region of residence at diagnosis (incorporated into the random effects). No treatment information is available from the QCR.

The geographic regions used were 478 Statistical Local Areas (SLAs), defined under the Australian Bureau of Statistics' Australia Standard Geographic Classification (ASGC) [20]. Geocoded cancer patient residence information was assigned an SLA prior to data extraction using the 2006 ASGC boundary definitions. SLAs cover Queensland without gap or overlap, and in 2006 had populations ranging from 7 to 74,804 (median 5,723).

The expected hazard rate ($h^*(t)$ in (1)) was calculated from population and mortality data. Population data by 5-year age groups, sex, year and SLA were obtained from the Australian Bureau of Statistics [21]. Unit-record level mortality data for Queensland residents were obtained from the Australian Bureau of Statistics (to 2005) [22] and the Australian Coordinating Registry (2006-2011) [23]. The SLA boundaries provided in this mortality data changed over time. These were adjusted to the 2006 SLA boundaries using correspondence files produced by the Australian Bureau of Statistics. Population mortality estimates for each SLA, sex, integer age and year was calculated using sex-specific aggregated 5-year age groups, 5-year time periods (1997-2001, 2002-2006, 2007-2011), and over the same groups of SLAs used in the neighborhood structure for the CAR distribution. This smoothing enabled more stable population mortality estimates.

The remoteness of each SLA was assigned based on the Accessibility Remoteness Index of Australia plus (ARIA+), which has five categories ranging from "Major City" to "Very Remote", and is based on access to services and population sizes.

Three categories of tumor stage were included: Localized, Advanced and Unknown. Colorectal cancer stage was defined using the Dukes staging system after extracting information from pathology records held by the QCR [24]. Although four stage categories were defined, these were aggregated for increased accuracy [24]. 'Localized' stage cancer was defined as stages I and II, with 'Advanced' stage cancer defined as stages III and IV. Breast cancer stage was approximated based on information routinely collected by the QCR regarding tumor size, lymph node involvement and distant metastases, with 'Localized' stage equivalent to stage I, and 'Advanced' equivalent to stages II-IV [25]. No information was available on lung cancer stage. To assess the impact of stage at diagnosis on the survival differences between areas, separate models were run with and without the stage covariates.

Survival was calculated using the period method [26], with the 'at-risk' period covering 2002-2011. Under period analysis, all observations are left-truncated at the start of the at-risk period, in addition to being right-censored at the end [26]. This enables survival estimates to be based on more recent data, as for each cancer type included cases were the first primary cancer diagnosed during 1997 to 2011, and still alive during any part of 2002-2011. Data were censored at the 31st December 2011.

The analysis was conducted in two stages. First several versions of the non-spatial standard flexible parametric model were run using `stpm2` in Stata v13.1 (StataCorp LP, Texas, USA) to determine an appropriate transformation of the continuous variable patient age, as well as the preferred model form (hazards or odds) and number of pre-specified knots. These parameters were then used in the Bayesian spatial model, which was run with single chain MCMC using WinBUGS 1.4 (Imperial College and Medical Research Council, UK) interfaced with Stata. The first 250,000 iterations were discarded and a further 100,000 monitored (with every 10th iteration kept, for a total of 10,000). To enable the log likelihood (specified in (7)) to be calculated, what is referred to as the 'zeros trick' [27] was employed. WinBUGS code for the spatial flexible parametric relative survival model is supplied in Appendix A.

Age was included as a continuous variable by centering on the cancer-specific median age, then using fractional polynomial methods [28] in the non-spatial model to transform. A fractional polynomial extends a conventional polynomial by generalizing the powers to certain fractional and nonpositive values [29]. The Stata multivariable fractional polynomial

command was used (mfp), which fit different models with combinations of the default set of powers {-2, -1, -0.5, log, 0.5, 1, 2, 3} up to second-order fractional polynomials. For each cancer second-order fractional polynomials were preferred, although the transformations selected varied. Alternate nonlinear methods such as splines could have been used instead of fractional polynomials. Both are likely to give similar results, but the spline is more influenced by local variation as opposed to the global fractional polynomial [30].

The non-spatial model was also used to determine the appropriate number of pre-specified knots for the restricted cubic spline on the cumulative baseline hazard. Both BIC values and graphs of the estimated hazard and survival functions were used to select the preferred number of knots. The examination of graphs aimed to prevent overfitting models: if there were nominal differences between the plotted hazard or survival functions, then the model with the fewer number of knots was preferred.

A common output from a spatial analysis is a map of the estimates of interest. Under the Bayesian formulation, it is possible to map not only the median estimates of excess mortality odds ratios ($\exp(u_i + v_i)$), but also the probability of this ratio exceeding a certain value, such as 1. Thematic maps were produced using MapInfo Professional v12.5 (Pitney Bowes Software Inc., New York). Exceedance probability categories were defined based on standard 80% cut-offs, with a probability of 80% and above considered very likely to be true, while a probability below 20% is considered very unlikely to be true.

A key benefit in modelling a smooth baseline function is the ability to predict smooth survival or hazard functions. A range of post estimation commands are available for standard Royston-Parmar models in Stata. To obtain further benefit from the Bayesian approach, we derived additional syntax to calculate the predictions at each MCMC iteration (see Appendix B for Stata code). This enabled us to predict survival curves with appropriate credible intervals for each area. Predicted survival was calculated for each SLA, sex and cancer type at age 60 years, and further by stage (for breast and colorectal cancers).

4. Model Evaluation

Convergence of MCMC chains for each parameter was assessed using trace and density plots [27]. Due to the large number of areas, a subsample of 20 areas that included sparsely

populated areas (<0.2 residents per km^2) was selected for graphical monitoring of the u_i and v_i terms. No parameters showed evidence of non-convergence.

The accuracy of the posterior estimates was assessed using Monte Carlo (MC) error, calculated as $[\text{standard deviation}/\sqrt{\text{No. iterations}}]$ of the exponentiated odds ratio estimates for each parameter of interest (γ_d, β_k and $u_i + v_i$). As autocorrelation may influence MC error values, autocorrelation for each parameter was assessed via graphical plots, and generally found to be negligible, except in a few instances where the random effect estimation for some of the smaller regions took longer.

Sensitivity analyses [31] compared five different hyperpriors on the variance components σ_u^2 and σ_v^2:

Vague

1. Gamma distribution (shape, scale) on the precision, $\frac{1}{\sigma^2} \sim \Gamma(0.1,100)$
2. Uniform distribution (minimum, maximum) on the standard deviation, $\sigma \sim \text{U}(0.01,20)$
3. Uniform distribution on the standard deviation, $\sigma \sim \text{U}(0.01,100)$

Weakly informative

4. Uniform distribution on the standard deviation, $\sigma \sim \text{U}(0.01,2)$

Informative

5. Gamma distribution on the precision, $\frac{1}{\sigma^2} \sim \Gamma(1,2)$

Apart from version 5, similar estimates were obtained in each version. Examination of convergence trace and density plots for the four versions with comparable results indicated a slight preference for version 2, and for this reason results are presented based on version 2. Informative hyperpriors are expected to exert greater influence on the posterior, and version 5 imposes very low probability on the standard deviation being close to 0, so estimates had more variation than under the other hyperpriors considered. The associated supplementary material provides further details on hyperprior comparisons and a subsample of plots.

5. Results

Based on BIC under the non-spatial model, the proportional odds formulation was preferred over the proportional hazards form for the three cancers examined, except for breast cancer

unadjusted for stage, where the hazards form was marginally preferred (Table 1). Results presented here use the proportional odds form.

Although 4 interior knots were preferred for colorectal cancer based on BIC values (Table 1), graphs of the hazard function suggested fewer knots would suffice. The final number of interior knots selected was 2, 1 and 3 for breast, colorectal and lung cancers, respectively.

The flexible parametric form fitted the data better than using a piecewise approach or standard log-logistic distribution (Figure 1). Small numbers are likely to influence the rapid increase in mortality in the smoothed hazard function as time approaches 15 years after diagnosis. Although only shown for breast cancer, colorectal and lung cancer also exhibited similar patterns.

All models had very low MC error estimates. The maximum MC errors for any parameter were 0.0042, 0.0028 and 0.0029 for breast, colorectal and lung cancer, respectively (in the models unadjusted for tumor stage), and 0.0406 for breast and 0.0031 for colorectal cancer (in the models adjusted for tumor stage). MC errors within 5% of the parameter's standard deviation are considered acceptable [32], and all met this criteria.

All three cancers showed strong evidence of spatial inequalities in cancer survival after adjusting for age and sex (Figure 2). There was a consistent pattern of lower survival among remote areas, and higher survival among areas in the urban south-east corner. The probability of excess mortality odds ratios exceeding 1 was most definitive for lung cancer (Figure 2). This is partly influenced by the number of deaths, with more deaths providing greater precision.

After further adjusting for stage, most breast cancer median excess mortality odds ratios were somewhat attenuated, and this was most clearly demonstrated by the marked reduction in the probability of estimates differing from 1, with fewer remote areas showing high (>80%) probability of excess mortality odds ratios above the Queensland average (Figure 3). In contrast, colorectal cancer results were much less impacted by adjustment for tumor stage (Figure 3).

The median posterior predicted survival for each SLA was grouped by remoteness and cancer type to consider the range of survival predictions (Figure 4). Although urban areas often had

higher survival than remote areas, there was variation even within these groupings, and often the highest survival in remote areas was on par with the predicted survival in certain urban areas.

Survival curves were also predicted for each SLA. To illustrate the maximum survival differential, the major city SLA with the highest survival is compared against the very remote SLA with lowest survival (Figure 5). Survival differences for breast and colorectal cancer increased over time, while for lung cancer the largest inequalities were observed around 2-4 years after diagnosis, and these had diminished by 15 years (Figure 5). This is likely to reflect the very aggressive nature of lung cancer. The high number of deaths from lung cancer is also apparent by less uncertainty around the lung cancer survival estimates.

When assessing patterns by spread of disease, there were only small differences in survival observed for localized breast cancers, but much poorer survival for advanced breast cancers in the remote SLA compared to the major city SLA (Figure 5). The maximum absolute differential in 5-year survival between SLAs for advanced stage breast cancers was 6.7%, compared to 1.3% for localized breast cancers. In contrast, colorectal cancer showed marked survival differences between SLAs for localized cancers (maximum 5-year survival difference of 4.7%) and even higher for advanced cancers (14.0%).

6. Discussion

We have proposed an extension to Nelson's flexible parametric relative survival model to produce small-area survival estimates. This new model includes additional random effect components within a Bayesian framework to enable small-area smoothing. This combines the benefits of flexible parametric models: the smooth, well-fitting baseline hazard functions and predictive ability, with the Bayesian benefits of robust and reliable small-area estimates.

The predictive ability of these flexible parametric models remains an important advantage over piecewise based approaches. Concepts that are relative to the average, such as the excess mortality odds ratio, do not provide direct information on the survival impact, which is of most interest to cancer patients. Quantifying survival differences provides a more intuitive and balanced measure of inequalities, and we are not aware of survival curves being produced for such small areas previously.

Wide variation between SLAs was observed for predicted cancer survival, and this was particularly noticeable for cancers diagnosed at an advanced stage. Unlike breast cancer, comparatively large survival differences were observed even for colorectal cancers diagnosed at a localized stage. This is consistent with an impact of geographical differences in the management that patients receive depending on where they live [33]. Treatment data is not routinely collected by registries, so in the absence of data this remains speculative. However, multiple studies suggest that colorectal cancer patients often have better outcomes when treated by specialist surgeons with higher case volumes [34, 35]. In Queensland, these specialist surgeons with high-throughput are predominantly located in the urbanized south-east corner, which is classified as a major city region.

Robust survival estimates in small areas are only possible by incorporating spatial smoothing methods. Here, Bayesian methods used priors designed to smooth across neighboring regions, which produced reliable estimates and predictions despite data sparseness. As per the popular Besag, York and Mollie (BYM) model, two random effect area-level components were included with different priors: an intrinsic CAR normal prior for local smoothing and a normal distribution for global smoothing towards the overall mean [15]. Although the BYM model has been shown to perform well when compared to other Bayesian disease mapping approaches,[36] concerns have been raised about the potential for oversmoothing [37]. Investigating alternative approaches, such as including a component that allows for discrete changes between areas [38], could be a fruitful area for future research. This may be particularly important if the cancers of interest are strongly linked with known infectious or lifestyle components with large socioeconomic differentials, as neighboring regions can have markedly different socioeconomic profiles.

Although demonstrated here on relative survival, this model can easily be adjusted to model cause-specific or all-cause survival, broadening the diseases it can be applied to (see Appendix C for syntax). To interpret relative survival as net survival, conditional independence between mortality from the disease, and mortality due to other causes must be assumed [39]. For lung cancer, due to the association with smoking, this assumption is questionable [9]. However, due to the high mortality rate, the bias is small in practice [40]. One advantage then is the ease of also running a cause-specific analysis within the same modelling framework and comparing results.

Royston-Parmar models can be built on a range of parametric models. For all cancers

examined, the odds form was either preferable or similar to hazards. Other alternative model formulations include probit. Both odds and probit formulations assume non-proportional hazards that will converge to 1 as $t \to \infty$, although the probit form has slightly longer tails. Prognostic influences on cancer often demonstrate a diminishing impact on mortality as time from diagnosis increases [9].

Additional complexity could be incorporated into the models. In the current form of the model, the intercept contains the random effect spatial structure. This could be extended by incorporating random effects into the spline coefficients. Including additional temporal components could enable comparison of small-area variation in survival across time, or time-varying components could also be incorporated in a straightforward manner. These time-dependent effects can also be modelled using splines, providing a smooth, continuous hazard. Although in theory this is simple, the computational implementation may be challenging. Model averaging over different numbers of knots could be investigated, although given the similarity of functions once the number of knots reaches a certain threshold, unlikely to influence results. Allowing the number of knots to be determined within the Bayesian model would also be possible, but the insight into the model behavior and sensitivity obtained from comparing different numbers of knots was advantageous.

Perhaps the greatest disadvantage is the computational intensity of using MCMC analysis for these models. There is some theoretical justification for a long run of a single chain [41], and although we discarded 250,000 iterations in the results presented, convergence was achieved at far fewer iterations for most parameters. Although the time to run the entire 350,000 iterations was only slightly longer than the previous piecewise MCMC-based approaches used for small-area analyses (~8 hours on a high quality computer), even using a reduced number of iterations it is substantially longer than producing estimates analytically in Stata or R. Also, as the number of knots increased, computational time further increased. Using an MCMC approximation method such as INLA (Integrated Nested Laplace Approximation) [42] could overcome this difficulty.

In conclusion, these flexible parametric survival models have great potential for exploring spatial and spatio-temporal survival inequalities, and we hope to stimulate further development and application of these models within wider contexts.

Appendix A: WinBUGS code

WinBUGS code for the breast cancer spatial flexible parametric relative survival model including age (as a centered, continuous, transformed variable) and tumor stage. The 'zeros trick' is used to specify a general likelihood. By setting the observed data to a set of zeros, then a Poisson(c[i]) observation of 0 has a likelihood of exp(-c[i]), and here c[i]= - log(L[i]) (with a constant included to ensure the value is positive).

For further details on the components of the likelihood expression see Equation 7.

Note that haz=$h^*(t_i)$ = population mortality for each individual's age group, sex and year , N=number of data rows (individual-level observations), Nsla=number of areas, and sumNumNeigh=the overall sum of each area's number of neighbors.

Input data (in addition to the above) are:

rcs=restricted cubic spline terms (the associated number = number of internal knots minus one). This is computed externally in Stata prior to running in WinBUGS (as are s0rcs and drcs).
s0rcs= restricted cubic spline terms (delayed entry)
drcs=first derivative of restricted cubic spline terms
agec=transformed continuous, centered age values
stage=values representing tumor stage
slano=area id number
d=death indicator variable (0=censored, 1=death)
t0=late entry indicator variable (0=no late entry, 1=late entry)
t=time survived until death or censoring

```
model {
        K <-10000

        for(i in 1:N) {
           stage2[i]<- equals(stage[i], 2)
           stage3[i]<- equals(stage[i], 3)

        zeros[i]<-0
        eta[i]<-gamma[1]+gamma[2]*rcs1[i]+gamma[3]*rcs2[i]+gamma[4]*rcs3[i]+beta[1]*agec1[i]
        +beta[2]*agec2[i]+beta[3]*stage2[i]+beta[4]*stage3[i]+u[slano[i]]+v[slano[i]]

        eta0[i]<-gamma[1]+gamma[2]*s0rcs1[i]+gamma[3]*s0rcs2[i]+gamma[4]*s0rcs3[i]+beta[1]*agec1[i]
        +beta[2]*agec2[i]+beta[3]*stage2[i]+beta[4]*stage3[i]+u[slano[i]]+v[slano[i]]

        dsp[i]<-gamma[2]*drcs1[i]+gamma[3]*drcs2[i]+gamma[4]*drcs3[i]

        lnL[i]<- d[i]*log((haz[i])+(1/t[i])*max(dsp[i]*exp(eta[i]),0.00001)/(1 + exp(eta[i]))) +
                log(pow((1+exp(eta[i])),-1)) + (log(1+exp(eta0[i]))*t0[i])
        c[i]<- -lnL[i]+K
```

```
        zeros[i]~dpois(c[i])
        }

#Prior Distributions
#CAR prior for spatial random effect
        u[1:Nsla] ~ car.normal(adj[], weights[], num[], tauu)
        for (k in 1:sumNumNeigh) {weights[k] <- 1 .

#Normal prior for uncorrelated heterogeneity term
        for (i in 1:Nsla) {
        v[i]~dnorm(0,tauv)
        }

        # Other priors
        tauu<- pow(sigmau,-2)
        tauv<- pow(sigmav,-2)
        sigmau~dunif(0.01,20)
        sigmav~dunif(0.01,20)
        varucon <-1/tauu
        varv<-1/tauv
        varumarginal<-sd(u[])*sd(u[])
        fracspatial<-varumarginal/(varumarginal+varv)

for(j in 1:4){
beta[j]~dnorm(0,0.001)
}
for(j in 1:4){
gamma[j]~dnorm(0,0.001)
}

}
```

Appendix B: Survival calculations

Stata syntax for predicted survival calculations

```
*Breast cancer by stage survival predictions at age 60 years
set more off
*Calculate transformed age values at age 60 years (as age is centered, median=0 but represents 59 years)
  scalar a1=(((1+39)/10)^.5)-1.978965433
  scalar a2=(((1+39)/10)^2)-15.33743848
 scalar list
*Loop over each area
```

```
forvalues i=1/478{
*Skipped steps to read in and organise data, but have 10,000 rows with results from WinBUGS (gamma, beta, u, v) and input data to WinBUGS from Stata (rcs1, rcs2) for ~50 time points, producing a total of ~500,000 rows of data.
*Calculate predictions
* Loop over to generate results by stage, coded here as 0=localized, 1=advanced
forvalues s=0/1{
        preserve
        gen stage=`s'
        *Log odds of the probability of an event
        gen double h_`i'=gamma_1+(gamma_2*rcs1)+(gamma_3*rcs2)+(beta_1*a1)+(beta_2*a2)
        +(beta_3*stage)+u_`i'+v_`i'
        *Survival function (odds formulation)
        gen double s_`i'=(1+(exp(h_`i')))^-1
        * Calculate median and 80% credible interval values
        collapse (p50) h50_`i'=h_`i' s50_`i'=s_`i' (p10) h10_`i'=h_`i' s10_`i'=s_`i' (p90) h90_`i'=h_`i'
        s90_`i'=s_`i', by(_t)
        save _sf`s'_`i', replace
        list _t s* if _n==_N
        restore
        }
}
```

Appendix C: Modeling alternate survival estimates

The following shows the key Stata syntax and WinBUGS code to obtain alternative types of flexible parametric survival estimates.

It assumes the death variable is coded as 1 for the specific cancer death, 2 for all other deaths.

Relative survival

Stata syntax

```
*Include all deaths from any cause
* Use period approach
stset exit, enter(time mdy(1,1,2002)) exit(time mdy(12,31,2011)) ///
```

```
    origin(dxdate) failure(death==1,2) id(id) scale(365.24)

*Run the model
xi:mfp, select(0.05): stpm2 agec i.sex i.stage, df(2) scale(odds) bhazard(haz) nolog
```

WinBUGS code

```
*The log-likelihood
lnL[i]<- d[i]*log((haz[i])+(1/t[i])*max(dsp[i]*exp(eta[i]),0.00001)/(1 + exp(eta[i]))) +
log(pow((1+exp(eta[i])),-1)) + (log(1+exp(eta0[i]))*t0[i])
```

Cause-specific survival

Stata syntax

```
*Only include deaths attributed to the specific cancer
* Use period approach
stset exit, enter(time mdy(1,1,2002)) exit(time mdy(12,31,2011)) ///
    origin(dxdate) failure(death==1) id(id) scale(365.24)

*Run the model.
xi:mfp, select(0.05): stpm2 agec i.sex i.stage, df(2) scale(odds) nolog
```

WinBUGS code

```
*The log-likelihood
lnL[i]<- d[i]*log(max(dsp[i]*exp(eta[i]),0.00001)/(1 + exp(eta[i]))) +
log(pow((1+exp(eta[i])),-1)) + (log(1+exp(eta0[i]))*t0[i])
```

Overall survival (This is not an estimate of net survival)

Stata syntax

```
*Include all deaths from any cause
* Use period approach
stset exit, enter(time mdy(1,1,2002)) exit(time mdy(12,31,2011)) ///
    origin(dxdate) failure(death==1,2) id(id) scale(365.24)
```

The Stata model and WinBUGS log-likelihood is identical to cause-specific survival.

References

1. Lyseen AK, Nohr C, Sorensen EM *et al.* A review and framework for categorizing current research and development in health related Geographical Information Systems (GIS) studies. *Yearbook of Medical Informatics* 2014; **9**(1): 110-24.
2. Cramb SM, Mengersen KL, Baade PD. Developing the Atlas of Cancer in Queensland: Methodological Issues. *International Journal of Health Geographics* 2011; **10**: 9.
3. Yu XQ, O'Connell DL, Gibberd RW *et al.* Estimating regional variation in cancer survival: a tool for improving cancer care. *Cancer Causes & Control* 2004; **15**(6): 611-8.
4. Dickman PW, Sloggett A, Hills M *et al.* Regression models for relative survival. *Statistics in Medicine* 2004; **23**(1): 51-64.
5. Fairley L, Forman D, West R *et al.* Spatial variation in prostate cancer survival in the Northern and Yorkshire region of England using Bayesian relative survival smoothing. *British Journal of Cancer* 2008; **99**(11): 1786-93.
6. Cramb SM, Mengersen KL, Baade PD. *Atlas of Cancer in Queensland: geographical variation in incidence and survival, 1998 to 2007*. 2011, Viertel Centre for Research in Cancer Control, Cancer Council Queensland: Brisbane.
7. Saez M, Barceló MA, Martos C *et al.* Spatial variability in relative survival from female breast cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2012; **175**(1): 107-134.
8. Hennerfeind A, Held L, Sauleau EA. A Bayesian analysis of relative cancer survival with geoadditive models. *Statistical Modelling* 2008; **8**(2): 117-139.
9. Royston P, Lambert PC. *Flexible parametric survival analysis using Stata: beyond the Cox model*. 2011, College Station, Texas: StataCorp LP.
10. Nelson CP, Lambert PC, Squire IB *et al.* Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 2007; **26**(30): 5486-98.
11. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; **8**(5): 551-561.
12. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009; **9**(2): 265-290.
13. Jackson CH. *Package 'flexsurv', version 0.6*. 2015.
14. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* 2007, Cambridge: Cambridge University Press.
15. Besag J, York J, Mollie A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; **43**: 1-59.
16. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**(3): 515-533.
17. Queensland Cancer Registry. *Cancer in Queensland: Incidence, Mortality, Survival and Prevalence,1982 to 2011*. 2013, QCR, Cancer Council Queensland and Queensland Health: Brisbane.
18. Dasgupta P, Baade PD, Aitken JF *et al.* Multilevel determinants of breast cancer survival: association with geographic remoteness and area-level socioeconomic disadvantage. *Breast Cancer Research & Treatment* 2012; **132**(2): 701-10.
19. Baade PD, Dasgupta P, Aitken JF *et al.* Geographic remoteness, area-level socioeconomic disadvantage and inequalities in colorectal cancer survival in Queensland: a multilevel analysis. *BMC Cancer* 2013; **13**: 493.
20. Australian Bureau of Statistics. *Australian Standard Geographic Classification (ASGC), 2006*, in *ABS Cat. No. 1216.0*. 2006, ABS: Canberra.
21. Australian Bureau of Statistics. *Estimated Resident Population for QLD SLAs by 5 year age group and sex from 2001 to 2011 (based on ASGC 2006), Customised report*. 2013, Canberra: Regional Population Unit, ABS.

22. Australian Bureau of Statistics. *Unit record mortality data for Queensland by State of usual residence, 1982-2005*. (unpublished data). 2007, Canberra: ABS.
23. Registries of Births, Deaths and Marriages, the Coroners and the National Coronial Information System. *Unit record mortality data for Australia by State of usual residence, 2006-2011*. (unpublished data). 2014, Brisbane: Australian Coordinating Registry.
24. Krnjacki LJ, Baade PD, Lynch BM *et al.* Reliability of collecting colorectal cancer stage information from pathology reports and general practitioners in Queensland. *Australian & New Zealand Journal of Public Health* 2008; **32**(4): 378-82.
25. Cramb SM, Mengersen KL, Turrell G *et al.* Spatial inequalities in colorectal and breast cancer survival: Premature deaths and associated factors. *Health & Place* 2012; **18**(6): 1412-21.
26. Brenner H, Gefeller O, Hakulinen T. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computation realisation and applications. *European Journal of Cancer* 2004; **40**: 326-335.
27. Spiegelhalter DJ, Thomas A, Best N *et al. WinBUGS User Manual, version 1.4*. 2003, MRC Biostatistics Unit, Institute of Public Health: Cambridge.
28. Sauerbrei W, Meier-Hirmer C, Benner A *et al.* Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics & Data Analysis* 2006; **50**(12): 3464-3485.
29. Sauerbrei W, Royston P. Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1999; **162**(1): 71-94.
30. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine* 2013; **32**(13): 2262-2277.
31. Lambert PC, Sutton AJ, Burton PR *et al.* How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* 2005; **24**(15): 2401-28.
32. Powers DA, Xie Y. *Statistical Methods for Categorical Data Analysis, 2nd Edition*. 2008: Emerald.
33. Yu XQ, O'Connell DL, Gibberd RW *et al.* A population-based study from New South Wales, Australia 1996-2001: area variation in survival from colorectal cancer. *European Journal of Cancer* 2005; **41**: 2715-21.
34. Meagher AP. Colorectal cancer: is the surgeon a prognostic factor? A systematic review. *Medical Journal of Australia* 1999; **171**(6): 308-10.
35. Anwar S, Fraser S, Hill J. Surgical specialization and training - its relation to clinical outcome for colorectal cancer surgery. *Journal of Evaluation in Clinical Practice* 2012; **18**(1): 5-11.
36. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 2005; **14**(1): 35-59.
37. Goovaerts P, Gebreab S. How does Poisson kriging compare to the popular BYM model for mapping disease risks? *International Journal of Health Geographics* 2008; **7**: 6.
38. Lawson AB, Clark A. Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine* 2002; **21**(3): 359-70.
39. Gamel JW, Vogel RL. Non-parametric comparison of relative versus cause-specific survival in Surveillance, Epidemiology and End Results (SEER) programme breast cancer patients. *Statistical Methods in Medical Research* 2001; **10**(5): 339-52.
40. Hinchliffe SR, Rutherford MJ, Crowther MJ *et al.* Should relative survival be used with lung cancer data? *British Journal of Cancer* 2012; **106**(11): 1854-9.
41. Raftery AE, Lewis SM. [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. 1992: 493-497.
42. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B (Methodological)* 2009; **71**: 319-392.

Table 1: Variations in BIC values by model form and number of internal knots

		Unadjusted for tumour stage			Adjusted for stage	
	# knots	**Breast**	**Colorectal**	**Lung**	**Breast**	**Colorectal**
PH						
	0	43280	72198	42385	41227	67116
	1	43284	71747	40391	41234	66727
	2	43231*	71702	40294	41170	66667
	3	43237	71690	40175	41178	66654
	4	43244	71669	40183	41187	66624
	5	43251	71673	40197	41195	66625
	6	43261	71681	40201	41204	66636
	7	43265	71686	40195	41206	66639
PO						
	0	43286	71942	40431	41055	66754
	1	43294	**71698**	40300	41059	**66627**
	2	**43245**	71654	40281	**41004***	66580
	3	43251	71632	**40112***	41013	66555
	4	43259	71612*	40113	41021	66525*
	5	43266	71617	40127	41030	66530
	6	43276	71625	40135	41038	66539
	7	43280	71630	40135	41040	66543

BIC=Bayesian Information Criterion; PH=Proportional Hazards; PO=Proportional Odds

Notes: An asterisk denotes the lowest BIC value for each model, bolded values are the selected choice.

BIC values are comparing the non-spatial model versions in Stata.