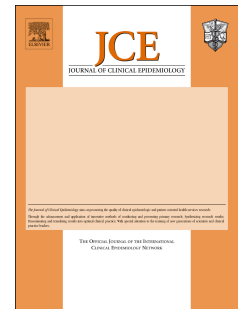# Journal Pre-proof

Uptake of methodological advances for synthesis of continuous and time-to-event outcomes would maximise use of the evidence base

Suzanne C. Freeman, Alex J. Sutton, Nicola J. Cooper

Please cite this article as: Freeman SC, Sutton AJ, Cooper NJ, Uptake of methodological advances for synthesis of continuous and time-to-event outcomes would maximise use of the evidence base, *Journal of Clinical Epidemiology* (2020), doi: https://doi.org/10.1016/j.jclinepi.2020.05.010.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Suzanne Freeman: Conceptualization, methodology, software, validation, investigation, data curation, writing – original draft

Alex Sutton: Conceptualization, methodology, validation, writing – review & editing

Nicola Cooper: Conceptualization, methodology, validation, writing – review & editing

Title: Uptake of methodological advances for synthesis of continuous and time-to-event outcomes would maximise use of the evidence base

Suzanne C Freeman[1], Alex J Sutton[1], Nicola J Cooper[1]

[1] Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK

Corresponding author:
Suzanne C Freeman, PhD
Department of Health Sciences
University of Leicester
Leicester, LE1 7RH
UK
suzanne.freeman@leicester.ac.uk

Running title: Uptake of methodological advances for synthesis of continuous and time-to-event outcomes

Abstract

Objective: To establish how often continuous and time-to-event outcomes are synthesised in health technology assessment (HTA), the statistical methods and software used in their analysis and how often evidence synthesis informs decision models.

Study design and setting: Review of NIHR HTA reports, NICE technology appraisals and NICE guidelines reporting quantitative meta-analysis or network meta-analysis of at least one continuous or time-to-event outcome published 01/04/2018 - 31/03/2019.

Results: We identified 47 eligible articles. At least one continuous or time-to-event outcome was synthesised in 51% and 55% of articles respectively. Evidence synthesis results informed decision models in two-thirds of articles. The review and expert knowledge identified five areas where methodology is available for improving the synthesis of continuous and time-to-event outcomes: i) outcomes reported on multiple scales, ii) reporting of multiple related outcomes, iii) appropriateness of the additive scale, iv) reporting of multiple time points and v) non-proportional hazards. We identified three anticipated barriers to the uptake and implementation of these methods: i) statistical expertise, ii) software and iii) reporting of trials.

Conclusion: Continuous and time-to-event outcomes are routinely reported in HTA. However, increased uptake of methodological advances could maximise the evidence base used to inform the decision making process.

Keywords: continuous outcomes, health technology assessment, time-to-event outcomes, evidence synthesis, decision models, clinical decision-making

What is new?

- Through a review of NIHR Health Technology Assessment reports, NICE technology appraisals and NICE guidelines published between 1st April 2018 and 31st March 2019 we establish that continuous and time-to-event outcomes are routinely synthesised within HTA articles. However, only two-thirds of articles used the results from an evidence synthesis of continuous or time-to-event outcomes to inform an economic decision model.

- Evidence from the review combined with expert knowledge of the methodological field identified five key areas where under-utilised methodology is available for improving the synthesis of continuous and time-to-event outcomes: i) outcomes reported on multiple scales, ii) reporting of multiple related outcomes, iii) appropriateness of the additive scale, iv) reporting of multiple time points and v) non-proportional hazards.

- We identified three anticipated barriers to the uptake and implementation of these methods: i) availability of specialist statistical expertise for model selection, model fitting and interpretation of results, ii) user-friendly software for implementing complex statistical and/or non-standard models, and iii) limited reporting of individual trials.

- More research is needed to develop, refine and generalise where possible methods for synthesising continuous and time-to-event outcomes for the purpose of decision modelling to maximise the evidence base utilised in the decision making process.

1. Introduction

Health Technology Assessment (HTA) is a form of policy research that examines the short- and long-term consequences related to the use of a health technology. The purpose of HTA is to facilitate decision making by providing information about the most clinically effective and cost-effective interventions for a given condition, and to provide this information in a systematic, transparent, and unbiased manner. Evaluating health technologies using this evidence-based approach is a key component of healthcare decision making often resulting in cost savings and better quality of treatment for patients [1].

Evidence synthesis is a well-established component of HTA applied to quantitatively combine the data from multiple trials in order to obtain an overall pooled estimate(s) of clinical effectiveness, which may be used to inform an associated economic evaluation. For comparisons between two healthcare interventions, it is common practice to apply pairwise meta-analysis (MA) methods to obtain pooled effectiveness estimates; however, where more than two interventions are of interest network meta-analysis (NMA) [2] (also known as multiple treatment comparisons [3] or mixed treatment comparisons [4]) is applied. NMA extends pairwise MA to allow the simultaneous estimation of comparative effectiveness of multiple interventions using an evidence base of trials that individually may not compare all intervention options, but together form a connected network of comparisons.

Since 2004, in the UK, the National Institute for Health and Care Excellence (NICE) has recommended meta-analysis of randomised controlled trials (RCTs) as their preferred method for evidence synthesis[5]. In 2013 this recommendation was updated to acknowledge the role of NMA methods for assessing clinical effectiveness from all relevant studies reporting clinically relevant outcomes [6]. It is a combination of the evidence on clinical effectiveness from evidence synthesis along with economic decision models which form the basis of NICE guidance for improving health and social care in the UK[1].

The synthesis of continuous and time-to-event outcomes is often perceived to be more complex than the synthesis of binary outcomes due to heterogeneous reporting of outcomes across trials, a higher propensity for outcomes to be missing, and difficult interpretability of results. In this paper we start by conducting a review of National Institute of Health Research (NIHR) HTA reports, NICE technology appraisals and NICE guidelines to establish the current state of play with regards to how often continuous and time-to-event outcomes are reported in HTA, the statistical methods and software used in practice for the synthesis of continuous and time-to-event outcomes, and how much of the evidence from evidence syntheses of continuous and time-to-event outcomes contributes to the economic decision model. From the review and expert methodological knowledge, we identify key challenges in synthesising continuous and time-to-event outcomes and provide examples of methodological advances to overcome these issues as well as identifying potential barriers to the implementation and uptake of these methods.

2. Review of NICE technology appraisals, NICE guidelines and NIHR HTA reports

2.1 Methods

We conducted a review of NICE technology appraisals, NICE guidelines and NIHR HTA reports published between the 1st April 2018 and the 31st March 2019. Articles were considered eligible for inclusion in this review if they contained a quantitative MA or NMA for at least one continuous or time-to-event outcome. Articles were excluded if they reported diagnostic outcomes, prognostic outcomes, feasibility studies or were updates of previous reviews without evidence synthesis. NICE technology appraisals and NICE guidelines were identified from the lists of published technology appraisals and guidelines on the NICE website (https://www.nice.org.uk/guidance/published?type=ta and https://www.nice.org.uk/guidance/published?type=apg,csg,cg,mpg,ph,sg,sc). NIHR HTA reports were identified from the NIHR HTA website (https://www.journalslibrary.nihr.ac.uk/hta/volume/?volume=22#/ and https://www.journalslibrary.nihr.ac.uk/hta/volume/?volume=23#/). Articles were screened for eligibility by one author (SCF). Articles where eligibility was unclear were discussed by all authors.

We extracted details relating to the outcomes and analysis methods for evidence synthesis of clinical data, details relating to the use of a decision model and how this may have been informed by an evidence synthesis as well as software. Informed by our expert methodological knowledge, we specifically extracted details relating to analysis of multiple outcomes, multiple time points and standardisation to determine whether recent advances in evidence synthesis methodology, to address these common issues, have filtered through to the HTA process. Items that were pre-specified for extraction are listed in Appendix A. Based on our knowledge of the new methodologies we identified what we anticipate to be three barriers to implementing these methods.

For NICE guidelines, which often consist of multiple evidence reviews, we included the first evidence review, which reported both an eligible evidence synthesis and an economic analysis. If none of the evidence reviews included an economic analysis then the first evidence review to consider an eligible evidence synthesis was included. Any guideline, which did not include a methods section for evidence synthesis, was excluded from this review.

For NICE technology appraisals, which often consist of multiple documents from committee meetings we only extracted data from the company submission document. Where this indicated the presence of additional appendices containing further details about the evidence synthesis methodology we requested the appendices from NICE. A total of 25 appendices were requested with 10 appendices received. Terminated appraisals were excluded from this review.

2.2 Results

Between 1st April 2018 and 31st March 2019, 56 NICE technology appraisals, 27 NICE guidelines and 69 NIHR HTA reports were published. A total of 47 articles were considered eligible for this review of which 25 (of 56, 45%) were NICE technology appraisals, 15 (of 27, 56%) NICE guidelines and 7 (of 69, 10%) NIHR HTA reports (Figure 1). A list of the articles included in this review can be found in Appendix B.

Continuous and time-to-event outcomes were routinely reported with 51% (24) of articles reviewed reporting at least one continuous outcome and 55% (26) of articles reporting at least one time-to-

event outcome. Three articles reported both continuous and time-to-event outcomes. The results for continuous outcomes are presented in Table 1 and for time-to-event outcomes in Table 2 with key points discussed below.

2.2.1 Continuous outcomes: Which methods of evidence synthesis are used in practice?

Pairwise MA were reported for all NICE guidelines and NIHR HTA reports. Whereas NMA was reported in all NICE technology appraisals. Across all reviews, the most common outcome measure (for the first reported continuous outcome) was the mean difference (MD). Alternative outcome measures included standardised mean difference (SMD), mean, percentage, risk difference and relative risk. Articles reporting continuous outcomes covered a wide range of clinical fields and included both pharmacological and non-pharmacological interventions. As such, a wide range of outcomes were considered including disease-specific scales (e.g. Alzheimer's disease assessment scale) and truly continuous measures (e.g. weight change, Appendix Table C.1).

Nine (69%) NICE guidelines reported multiple pairwise comparisons and one NIHR HTA report of pairwise MA lumped interventions together. Statistical methods to analyse multiple outcomes simultaneously were not reported in any articles. Multiple time points were reported in nine articles and were handled by reporting separate meta-analyses at specific time points. In total, only eight reviews (33%) used standardisation of any reported continuous outcomes.

All NICE guidelines and two NIHR HTA reports used Cochrane Review Manager (RevMan) to synthesise clinical evidence. Alternative software options included R, Stata and WinBUGS. Where reported, the most commonly used software for cost effectiveness was Microsoft Excel.

2.2.2 Time-to-event outcomes: Which methods of evidence synthesis are used in practice?

All NICE guidelines reported pairwise MA whereas NMA was more frequently reported in NICE technology appraisals and NIHR HTA reports. The majority of articles were pharmacological interventions in oncology. The most common outcomes were overall and progression-free survival (Appendix Table C.2). As such, the most common outcome measure for reporting time-to-event outcomes was the hazard ratio. Alternative outcome measures included time-varying hazard ratios, fractional polynomial coefficients, median survival, mean survival and relative risk.

Multiple pairwise comparisons were reported in eight articles (31%) and in one NICE technology appraisal, two drug interventions were lumped together for inclusion in the NMA alongside eight other interventions. Combining these two interventions into one treatment node for the NMA assumes that both interventions have the same efficacy across trials for the endpoints assessed. Multiple time points were reported in three articles and in all articles were handled by reporting separate meta-analyses at specific time points. No articles reported analysing multiple outcomes simultaneously.

Reporting of software used for evidence synthesis of clinical data was poor for NICE technology appraisals. As with continuous outcomes, all NICE guidelines reported using Cochrane Review

6

Manager (RevMan) to synthesise the clinical evidence and alternative software options included R, Stata and WinBUGS. Where reported, Microsoft Excel was used for cost-effectiveness analyses.

2.2.3 Continuous outcomes: How do the evidence synthesis results inform the economic decision models?

Decision models were reported in the majority of articles across all review types for continuous outcomes (17, 71%). The decision models were informed by clinical results in 16 reviews and of these 11 (69%) were informed by the evidence synthesis result. The decision models in the remaining five trials were informed by single trials only.

2.2.4 Time-to-event outcomes: How do the evidence synthesis results inform the economic decision models?

All articles reporting a time-to-event outcome reported a decision model, with all articles reporting that clinical results were used to inform the decision model. Overall, the decision models in most articles were informed by evidence synthesis results (17, 65%). However, half of the NICE guidelines, four NICE technology appraisals and two NIHR HTA reports were informed by a single trial despite an evidence synthesis being conducted. Reasons for only using a single trial ranged from article-specific considerations such as "only one study measured health related quality of life in UK patients with non-small-cell lung cancer" to unclear statements such as "one study identified as most suitable".

3. Challenges in synthesising continuous and time-to-event outcomes

In the review above MA were often conducted separately for each time point and/or outcome. This approach to analysis lacks coherence as a different evidence base is used in each analysis and restricts the usefulness of the evidence synthesis for informing the decision model. It would be better to extend the evidence synthesis methodology to allow for multiple outcomes and/or multiple time points (as well as multiple treatments). In this way, the evidence base used for decision making can be maximised. Below we highlight five of the key challenges for synthesising continuous and time-to-event outcomes and describe some of the methods available to overcome these challenges and maximise the evidence base available. We highlight key examples from the review itself and draw on the literature to promote and illustrate their potential for improving syntheses of continuous and time-to-event outcomes in the future.

3.1 Outcomes reported on multiple scales

For continuous outcomes, trials may individually measure an outcome on multiple instruments, while only one is chosen to contribute to the synthesis. This can mean that some outcome data is excluded from the MA [7]. To avoid excluding data in clinical areas, such as pain and anxiety, where outcomes are frequently measured across multiple scales they are often analysed as SMD. However, SMD are hard to interpret in a meaningful way and are limited in their ability to be included in decision models [8]. Methods to overcome the problems with SMD include transformation back to one of the original scales or conversion to relative risk [9].

7

In our review of NIHR HTA reports, NICE technology appraisals and NICE guidelines we identified one NICE guideline which reported a NMA of SMD and included the result of the NMA in the decision model [10]. NG116 reported an NMA of changes in post-traumatic stress disorder (PTSD) symptoms between baseline and treatment endpoint with alternative scenarios considering change in PTSD symptoms between baseline and 1-4 month follow-up. The decision model required probabilities of remission for each treatment which could not be directly estimated from the SMD. Therefore, SMD was transformed into a log odds ratio and then exponentiated into an odds ratio from which the probability of remission for each intervention was calculated. Converting SMD to a log odds ratio requires the assumption that there is an arbitrary cut off point on the underlying scale for determining whether a treatment is effective or not. The disadvantages of dichotomising continuous outcomes in this way have been well documented [11-13]. However, by making this assumption the authors were able to make use of all the available evidence.

When data are skewed methods for combining medians have been shown to perform better than SMD[14]. However, another alternative to standardisation is mapping to a common scale. This is a method of synthesis on ratios of different outcome scales [7,15,16] which allows data on all scales to be synthesised simultaneously and contribute to a single outcome expressed on each of the native scales. An advantage of this approach in a decision making scenario is that an appropriate scale for inclusion in the economic model can be easily chosen. Lu *et al*[16] fit these models in a Bayesian framework using WinBUGS. WinBUGS is a specialist statistical software package requiring specialist statistical expertise to ensure appropriate prior distributions are selected and the models parameters and results are interpreted correctly. Both the choice of software and the need for specialist statistical expertise are likely to be barriers to the uptake of this method. In our review of NIHR HTA reports, NICE technology appraisals and NICE guidelines, all NICE guidelines reported pairwise MA using RevMan to conduct their analyses. RevMan is limited in its analysis capabilities and uses a frequentist approach to analysis. To encourage the uptake of these methods within HTA articles it may be important to provide software training alongside training on the methods themselves.

3.2 Reporting of multiple related outcomes

For both continuous and time-to-event outcomes, multiple separate meta-analyses of related outcomes is undesirable because it limits the evidence base used and often hinders the exploration of important issues in the clinical area such as how intervention effects change over time and how different outcome scales relate to one another. This is particularly pertinent (although not exclusive) to the NMA setting where different interventions may have been evaluated using different outcome measures and thus choosing specific outcomes causes systematic exclusion.

Multivariate MA can allow the synthesis of multiple outcomes on their original scales and account for multiple correlated outcomes, which are often reported in RCTs, for example systolic and diastolic blood pressure, or progression-free survival and overall survival [17]. This avoids the need to exclude any outcome data and means that multiple outcomes can be synthesised on their original scales. Accounting for the correlation between outcomes can result in increased precision in treatment effect estimates [18,19]. However, multivariate MA models become increasingly difficult to fit as data on each scale becomes sparse and/or the number of outcome scales used increases [17,20,21].

Multivariate MA models can be extended to the NMA setting [22]. Barriers to the uptake of multivariate MA and NMA are likely to include the need for specialist statistical expertise for both the fitting of the model and the interpretation of the model results as well as the need for software beyond RevMan. A recently published NICE Technical Support Document on multivariate meta-analysis may help make this methodology more accessible to the HTA audience [23].

An NMA of antivirals for the treatment of influenza is an example of an NMA specifically developed for NICE to identify which antiviral was most effective at preventing influenza but for which the relevant trials reported different outcomes [24]. The network consisted of three antivirals each directly compared with placebo but without any trials directly comparing antivirals to each other. In the analysis of individual studies, studies reported either time to alleviation of fever, time to alleviation of symptoms or both outcomes. Synthesis of these studies was further inhibited by the reporting of different summary measures e.g. mean time, median time and proportion symptom free at the end of follow-up. To overcome these complexities, an NMA model consisting of a Weibull model with exchangeable treatment effects that were independent for each outcome but had a common random effect mean for the two outcomes was fitted [24]. The advantage of this approach was the simultaneous synthesis of two outcomes estimating a single summary measure allowing conclusions to be drawn about the comparative efficacy of four treatments for which head-to-head trials did not exist and preventing the need to either exclude relevant trials or conduct multiple MA. In a decision making scenario the advantage of this approach is that the single Weibull model parameter estimated can be easily incorporated within a decision model.

3.3 Appropriateness of the additive scale

Previous research for the synthesis of binary outcomes has focused on the appropriateness of different outcome scales (e.g. odds ratios vs. relative risks vs risk differences) [25]. However, for continuous outcomes additive scales have, almost unquestionably, been used (i.e. differences in means between groups). In recent MA of continuous outcomes, particularly in the areas of pain and depression [26-29], between-study heterogeneity has been explained (in part) by the inclusion of covariates to account for differences between trials in the severity of pain/depression at baseline. In this case, an additive scale and hence the use of MD and SMD may not be appropriate. An alternative outcome measure on a multiplicative scale, the ratio of means, has been proposed in a series of recent papers [30-32]. Using a multiplicative scale avoids the strong assumption that all tests are linear transformations of the same underlying measurement scale. In addition, the ratio of means does not require knowledge of the pooled standard deviation, which is required for SMD but is often unknown by clinicians, and therefore has a more natural clinical interpretation [30,31]. Furthermore, in contrast to the SMD, the ratio of means provides a direct probability that can easily be incorporated within a decision model. However, in our review of NIHR HTA reports, NICE technology appraisals and NICE guidelines we found no articles reporting the ratio of means. Despite being a fairly straightforward method to implement, requiring only the mean outcome from each treatment arm, there appears to be a lack of knowledge about the ratio of means, the assumptions it makes and when it may be appropriate to use. Uptake of the ratio of means may increase over the next few years following its addition to the Cochrane Handbook for Systematic Reviews of Interventions (Chapter 6.5.1.3) where it is recommended as an effect measure for outcomes which are physical measurements taking only positive values [33].

3.4 Multiple time points

For both continuous and time-to-event outcomes, appropriately dealing with multiple time points is especially important in the HTA setting, as a key component of decision models is modelling effectiveness over the full follow-up duration so it is essential that any trends over time are incorporated within the decision model. RCTs often report the same outcomes at multiple time points, for example pain at 2, 4 and 8 weeks post-operation. However, studies, which do not report the outcome of interest at the specific time point of interest, can be excluded from the synthesis. As a result, several methods to account for multiple time points have been proposed in the literature [34-36] with a recent review comparing three of these methods [37]. Methods for accounting for multiple time points depend heavily on the time points reported in the individual trials and are likely to require specialist statistical expertise for the interpretation of results and software beyond RevMan.

Lu *et al.*[38] accounted for outcomes reported at multiple time points using a piecewise exponential model. The network consisted of six treatments and 41 RCTs with RCTs reporting the outcome of interest, healing rate, at least once at 4, 6, 8 or 12 weeks. At each time point, the network diagram took on a different shape reflecting the variable evidence base available. An NMA model was fitted in which the time horizon was divided into four time periods with a piecewise hazard for each time period. A treatment effect was placed on time period 1 with treatment effects in subsequent time periods related to period 1 via a random walk model, which assumed that treatment effects were similar in each time period. The between-trial correlations between the treatment effects at different time points are implicit in the random walk variance. This is an example of simultaneously 'borrowing strength' on comparative effects across time periods while combining direct and indirect evidence on treatment comparisons within each time period. The model produced a hazard ratio for each treatment effect for each time period. This example demonstrates the potential for NMA methodology to account for multiple time points allowing the synthesis of a greater proportion of data from the evidence base whilst providing treatment effect estimates which can be included in decision models in a straightforward manner using techniques already widely in use.

In clinical areas such as oncology, epilepsy and progressive diseases RCTs often report both longitudinal and time-to-event outcomes. Methods have recently been developed that allow MA of joint longitudinal and time-to-event models [39]. Specific software packages for the fitting of joint models have been developed in both Stata [40] and R [41,42]. Specialist statistical expertise for the fitting of models and interpretation of results as well as specialist software are likely to be barriers to the uptake of joint models.

3.5 Non-proportional hazards

Conventionally time-to-event outcomes are analysed using a two-stage modelling approach in which the Cox model is used to obtain an estimate of the hazard ratio and its standard error for each trial in the first stage before synthesising them using a fixed effect or random effects MA model in the second stage. The Cox model assumes proportional hazards over time and it has been shown that this is not always a reasonable assumption, particularly in oncology when comparing chemotherapy and immunotherapy drugs [43]. In addition, clinicians can find hazard ratios hard to interpret [44].

Alternatives to the Cox model which can also allow for non-proportional hazards include fractional polynomial models [45], the piecewise exponential model [38,46], the Royston-Parmar model [47] and parametric survival curves [48]. Furthermore, a wider range of models can be fitted if a one-stage modelling approach is used. In the presence of non-proportional hazards, a single hazard ratio is no longer a suitable outcome measure. Alternative outcome measures, which have been used to synthesise evidence from multiple trials, include restricted mean survival time [49,50] and accelerated failure time [51-53]. The use of one-stage modelling approaches and alternative outcome measures are dependent on the data available from individual trials. Both one-stage models and restricted mean survival time require either individual participant data (IPD) or the reconstruction of IPD from Kaplan-Meier plots using methods such as the Guyot algorithm [54].

In our review of NIHR HTA reports, NICE technology appraisals and NICE guidelines the hazard ratio was the most commonly reported outcome measure for time-to-event outcomes. However, we also identified alternative outcome measures in almost a quarter of NICE technology appraisals, with the most popular method being fractional polynomial models. For example, in TA520 an NMA was conducted to assess the clinical effectiveness of atezolizumab compared to docetaxel and nintedanib to inform a cost-effectiveness analysis of atezolizumab. The proportional hazards assumption was tested and found to be violated so a fractional polynomial approach was used for the NMA. Barriers to the uptake and implementation of methods for dealing with non-proportional hazards are likely to include specialist statistical expertise for the fitting of models and interpretation of results as well as the use of software such as R, Stata or WinBUGS.

4. Implication of findings

From a combination of our review of NICE technology appraisals, NICE guidelines and NIHR HTA reports and our expert knowledge of the relevant methodological area, we have identified five key areas where methodology is available for improving the synthesis of continuous and time-to-event outcomes within HTA. In the process, we have also identified what we believe to be are three barriers to the uptake and implementation of those methods:

i)   Availability of specialist statistical expertise

Statistical expertise is advised for all MA and NMA; however, the methods described above often require specialist statistical expertise in advanced evidence synthesis and/or survival analysis methodology to aid appropriate model selection, model fitting and interpretation of results, which may not be available to all systematic review groups.

ii)   User-friendly software for implementing complex statistical and/or non-standard models

The complex statistical and/or non-standard models described above often require either specific statistical software (e.g. WinBUGS) or development of bespoke statistical routines in packages such as R or Stata. For example, in TA520 the clinical effectiveness of atezolizumab was assessed using the fractional polynomial NMA model (as described above). Cost-effectiveness was assessed using a three state (on treatment, off treatment and death) partitioned survival analysis model. The model was constructed by calculating the proportion of patients in each

11

health state based on time-to-treatment discontinuation and overall survival curves, and the difference between the two curves, at discrete time points. Comparator curves for docetaxel and nintedanib were constructed by using atezolizumab as the reference and applying the time-dependent hazard ratios from the fractional polynomial NMA model. Fractional polynomials can be fitted in Stata, R and WinBUGS but are not available in, the most commonly reported software used for synthesis, RevMan. Development of user-friendly software for implementing methods such as fractional polynomial NMA and other methods described in Section 3 may improve the uptake of these methods.

iii) Limited reporting of individual trials.

Limited reporting of outcomes data within individual trial publications/reports can restrict the choice of methods available for the analysis. For example, when using multivariate MA a more precise estimate of the correlation between outcomes can be obtained if some trials report both outcomes. For all MA a wider range of models are available if IPD is available for synthesis. Whilst we acknowledge that there is often a word limit for journal publications we encourage researchers to publish analyses not included in the main article in online appendices and encourage the use of data repositories where possible.

5. Conclusion

Continuous and time-to-event outcomes are routinely synthesised within HTA articles. However, only approximately two-thirds of articles used the results of evidence synthesis to inform the decision model. Through our review of NIHR HTA reports, NICE technology appraisals and NICE guidelines we have identified five key areas where methodology is available for improving the synthesis of continuous and time-to-event outcomes: i) outcomes reported on multiple scales, ii) reporting of multiple related outcomes, iii) appropriateness of the additive scale, iv) reporting of multiple time points and v) non-proportional hazards. In addition we have identified three potential barriers to the uptake and implementation of these methods: i) availability of specialist statistical expertise for model selection, model fitting and interpretation of results, ii) user-friendly software for implementing complex statistical and/or non-standard models, and iii) limited reporting of individual trials. Addressing these barriers could increase the evidence base used within evidence syntheses of continuous and time-to-event outcomes and maximise the evidence base utilised in the decision making process. Therefore, it is important that analysts and decision modellers involved in the HTA process are aware of the expanding literature for the synthesis of continuous and time-to-event outcomes and appreciate the limitations of simpler approaches. Nevertheless, more research is needed to develop, refine and generalise where possible methods for synthesising continuous and time-to-event outcomes in this context for the purpose of decision modelling.

References

1. Cooper N, Sutton A, Achana F, Welton N. Use of network meta-analysis to inform clinical parameters in economic evaluations. *Canadian Agency for Drugs and Technologies in Health.* 2016.
2. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med.* 2002;21(16):2313-2324.
3. Caldwell D, Ades A, Higgins J. Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *BMJ.* 2005;331:897-900.
4. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods.* 2012;3(2):80-97.
5. National Institute of Health and Care Excellence. Guide to the methods of technology appraisal 2004. Available at: https://webarchive.nationalarchives.gov.uk/20080611223138/http://www.nice.org.uk/niceMedia/pdf/TAP_Methods.pdf. Published 2004. Accessed.
6. National Institute of Health and Care Excellence. Guide to the methods of technology appraisal 2013. Available at: https://www.nice.org.uk/process/pmg9/chapter/foreword. Published 2013. Accessed.
7. Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. *Res Synth Methods.* 2015;6(1):96-107.
8. Lee YY, Le LK, Stockings EA, et al. Estimation of a Relative Risk Effect Size when Using Continuous Outcomes Data: An Application of Methods in the Prevention of Major Depression and Eating Disorders. *Med Decis Making.* 2018;38(7):866-880.
9. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ.* 2019;364:k4817.
10. NICE guideline NG116. *Post-traumatic stress disorder - Evidence reviews for psychological, psychosocial and other non-pharmacological interventions for the treatment of PTSD in children and young people.* 2018.
11. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ.* 2006;332:1080.
12. Senn S. Disappointing dichotomies. *Pharmaceutical Statistics.* 2003;2:239-240.
13. Ofuya M, Sauzet O, Peacock JL. Dichotomisation of a continuous outcome and effect on meta-analyses: illustration of the distributional approach using the outcome birthweight. *Syst Rev.* 2014;3:63.
14. McGrath S, Zhao X, Qin ZZ, Steele R, Benedetti A. One-sample aggregate data meta-analysis of medians. *Stat Med.* 2019;38(6):969-984.
15. Lu G, Brazier JE, Ades AE. Mapping from disease-specific to generic health-related quality-of-life scales: a common factor model. *Value Health.* 2013;16(1):177-184.
16. Lu G, Kounali D, Ades AE. Simultaneous multioutcome synthesis and mapping of treatment effects to a common scale. *Value Health.* 2014;17(2):280-287.
17. Bujkiewicz S, Thompson JR, Sutton AJ, et al. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Stat Med.* 2013;32(22):3926-3943.
18. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med.* 2011;30(20):2481-2498.
19. Riley RD, Jackson D, Salanti G, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ.* 2017;358:j3932.
20. Riley R, Abrams KR, Lambert PC. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat Med.* 2007;26:78-97.

21. Nam I, Mengersen K, Garthwaite P. Multivariate meta-analysis. *Stat Med.* 2003;22:2309-2333.

22. Jackson D, Bujkiewicz S, Law M, Riley RD, White IR. A matrix-based method of moments for fitting multivariate network meta-analysis models with multiple outcomes and random inconsistency effects. *Biometrics.* 2018;74(2):548-556.

23. Bujkiewicz S, Achana F, Papanikos T, Riley RD, Abrams KR. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. http://www.nicedsu.org.uk. Published 2019. Accessed.

24. Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. *Stat Med.* 2008;27(27):5620-5639.

25. Deeks JJ, Altman DG. *Effect measures for meta-analysis of trials with binary outcomes.* Second ed. London, UK: BMJ Publishing Group; 2001.

26. Freeman SC, Scott NW, Powell R, Johnston M, Sutton AJ, Cooper NJ. Component network meta-analysis identifies the most effective components of psychological preparation for adults undergoing surgery under general anaesthesia. *J Clin Epidemiol.* 2018;98:105-116.

27. Doleman B, Sutton AJ, Sherwin M, Lund JN, Williams JP. Baseline Morphine Consumption May Explain Between-Study Heterogeneity in Meta-analyses of Adjuvant Analgesics and Improve Precision and Accuracy of Effect Estimates. *Anesth Analg.* 2018;126(2):648-660.

28. Bower P, Konotopantelis E, Sutton AJ, et al. Influence of initial severity of depression on effectiveness of low intensity interventions: Meta-analysis of individual patient data. *BMJ.* 2013;346:f540.

29. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med.* 2008;5(2):e45.

30. Friedrich JO, Adhikari NK, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol.* 2011;64(5):556-564.

31. Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study. *BMC Med Res Methodol.* 2008;8:32.

32. Friedrich JO, Adhikari NK, Beyene J. Ratio of geometric means to analyze continuous outcomes in meta-analysis: Comparison to mean differences and ratio of arithmetic means using empiric data and simulation. *Stat Med.* 2012;31(17):1857-1886.

33. Higgins J, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions (2nd Edition).* Chichester, UK: John Wiley & Sons; 2019.

34. Dakin HA, Welton NJ, Ades AE, Collins S, Orme M, Kelly S. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Stat Med.* 2011;30(20):2511-2535.

35. Ding Y, Fu H. Bayesian indirect and mixed treatment comparisons across longitudinal time points. *Stat Med.* 2013;32(15):2613-2628.

36. Jansen JP, Vieira MC, Cope S. Network meta-analysis of longitudinal data using fractional polynomials. *Stat Med.* 2015;34(15):2294-2311.

37. Tallarita M, De Iorio M, Baio G. A comparative review of network meta-analysis models in longitudinal randomized controlled trial. *Stat Med.* 2019;38(16):3053-3072.

38. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med.* 2007;26:3681-3699.

39. Sudell M, Kolamunnage-Dona R, Gueyffier F, Tudur Smith C. Investigation of one-stage meta-analysis methods for joint longitudinal and time-to-event data through simulation and real data application. *Stat Med.* 2019;38(2):247-268.

40. Crowther MJ, Abrams KR, Lambert PC. Joint modeling of longitudinal and survival data. *The Stata Journal.* 2013;13(1):165-184.

41. Rizopoulos D. JM: An R Package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software.* 2010;35(9).

42. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. joineRML: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Med Res Methodol.* 2018;18:50.

43. Schadendorf D, Hodi FS, Robert C, et al. Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. *J Clin Oncol.* 2015;33(17):1889-1894.

44. Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. *Eur Heart J.* 2019;40(17):1378-1383.

45. Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol.* 2011;11:61.

46. Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Med Res Methodol.* 2012;12:34.

47. Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Res Synth Methods.* 2017;8(4):451-464.

48. Ouwens M, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Res Synth Methods.* 2010;1:258-271.

49. Wei Y, Royston P, Tierney JF, Parmar MKB. Meta-analysis of time-to-event outcomes from randomised trials using restricted mean survival time: Application to individual participant data. *Stat Med.* 2015;34:2881-2898.

50. Weir IR, Tian L, Trinquart L. Multivariate meta-analysis model for the difference in restricted mean survival times. *Biostatistics.* 2019.

51. Siannis F, Barrett JK, Farewell VT, Tierney JF. One-stage parametric meta-analysis of time-to-event outcomes. *Stat Med.* 2010;29:3030-3045.

52. Cope S, Jansen JP. Quantitative summaries of treatment effect estimates obtained with network meta-analysis of survival curves to inform decision-making. *BMC Medical Research Methodology.* 2013;13:147.

53. Barrett JK, Farewell VT, Siannis F, Tierney JF, Higgins J. Two-stage meta-analysis of time-to-event data from individual participants using percentile ratios. *Stat Med.* 2012;31:4296-4308.

54. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology.* 2012;12(1):9.

Figure 1: Flowchart of article selection process. HTA = health technology assessment, NICE = National Institute for Health and Care Excellence, NIHR = National Institute for Health Research
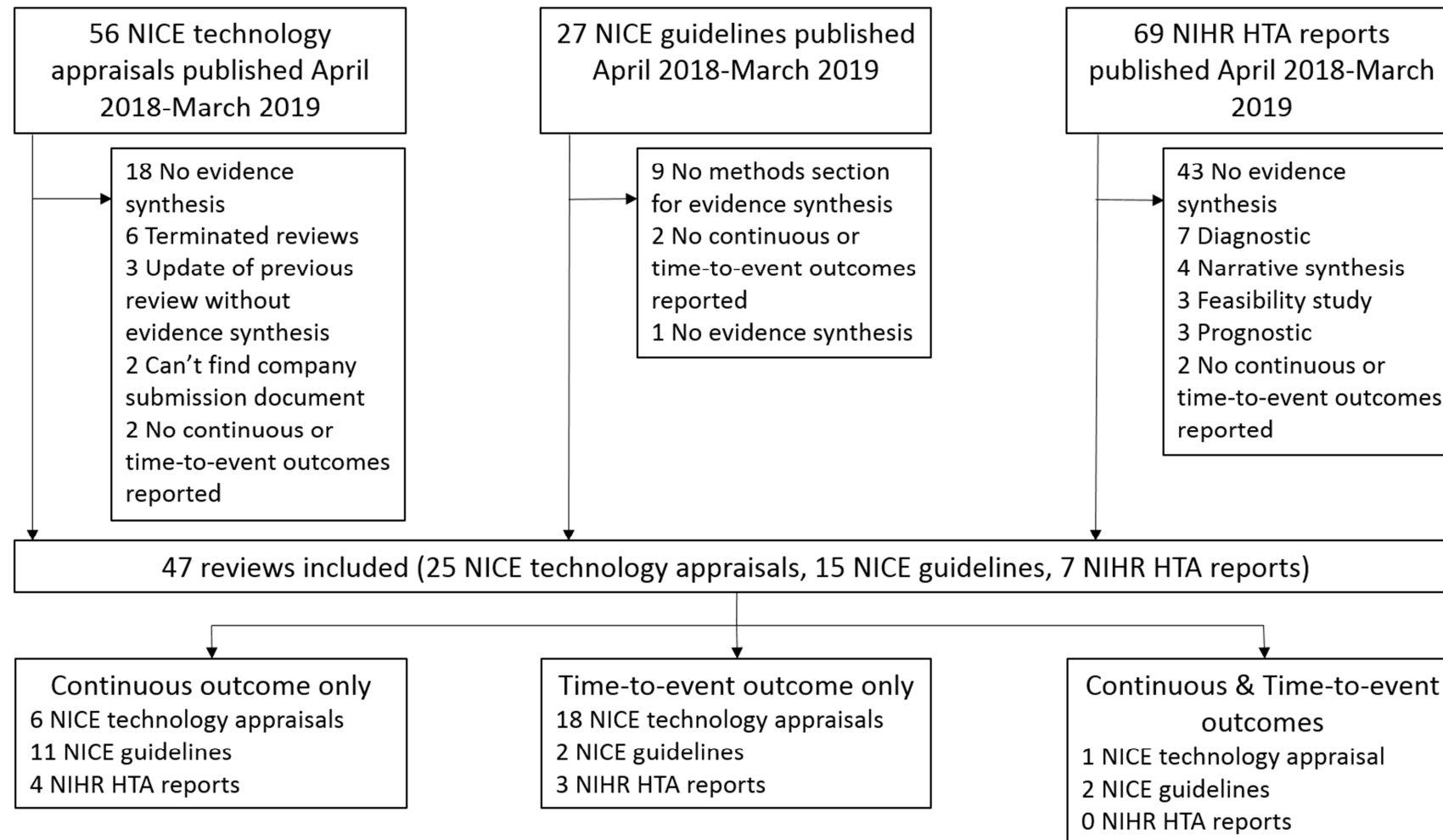
Table 1: Results for continuous outcomes

	NICE Technology Appraisal	NICE Guidelines	NIHR HTA Report
No. of articles	**7**	**13**	**4**
Clinical Field			
Cardiovascular	0	1 (8%)	1 (25%)
Dementia	0	1 (8%)	1 (25%)
Kidney	0	2 (15%)	0
Mental health	0	2 (15%)	0
Psoriatic arthritis	2 (29%)	0	0
Respiratory	1 (14%)	1 (8%)	0
Skin	2 (29%)	0	0
Cerebral palsy	0	1 (8%)	0
Diabetes	1 (14%)	0	0
Hearing loss	0	1 (8%)	0
Lyme disease	0	1 (8%)	0
Obesity	0	0	1 (25%)
Oncology	1 (14%)	0	0
Pancreatitis	0	1 (8%)	0
Pharmacy	0	1 (8%)	0
Rheumatoid arthritis	0	1 (8%)	0
Sleep	0	0	1 (25%)
Type of intervention			
Pharmacological	7 (100%)	4 (31%)	0
Non-Pharmacological	0	4 (31%)	2 (50%)
Both	0	5 (38%)	2 (50%)
SYNTHESIS			
Method of synthesis			
Meta-analysis	1 (14%)	13 (100%)	4 (100%)
Network meta-analysis	7 (100%)	3 (23%)	0
Outcome measure			
Mean difference	1 (14%)	11 (85%)	2 (50%)
Standardised mean difference	1 (14%)	1 (8%)	1 (25%)
Median difference	1 (14%)	0	0
Mean	1 (14%)	0	0
Percentage	2 (29%)	0	1 (25%)
Risk difference	1 (14%)	0	0
Relative risk	0	1 (8%)	0
Fixed effect or random effects			
Fixed effect	0	1 (8%)	0
Random effects	0	1 (8%)	4 (100%)
Both	4 (57%)	11 (85%)	0
Unclear	3 (43%)	0	0

17

Analysis details			
Multiple outcomes simultaneously	0	0	0
Lumping of interventions	0	0	1 (25%)
Multiple pairwise comparisons	1 (14%)	9 (69%)	1 (25%)
Standardisation of any outcomes	1 (14%)	5 (38%)	2 (50%)
Multiple time points	1 (14%)	7 (54%)	1 (25%)
Presentation of results			
Forest plot	1 (14%)	13 (100%)	3 (75%)
Network diagram (if NMA only)	4/7 (57%)	2/3 (67%)	0/0
Tables	6 (86%)	2 (15%)	3 (75%)
Other	1 (14%)	0	0
Software for synthesis			
Not reported	4 (57%)	0	1 (25%)
R & WinBUGS	1 (14%)	0	0
Cochrane Review Manager	0	10 (77%)	2 (50%)
Cochrane Review Manager & R	0	1 (8%)	0
Cochrane Review Manager & WinBUGS	0	2 (15%)	0
Stata	1 (14%)	0	1 (25%)
Stata & WinBUGS	1 (14%)	0	0
DECISION MODEL			
Decision model details			
Decision model	6 (86%)	8 (62%)	3 (75%)
Graph of model structure	5/6 (83%)	7/8 (88%)	1/3 (33%)
Clinical results inform decision model	6/6 (100%)	7/8 (88%)	3/3 (100%)
How much clinical evidence informs decision model?			
Meta-analysis	0	2/8 (25%)	2/3 (67%)
Network meta-analysis	5/6 (83%)	2/8 (25%)	0
Not reported	0	1/8 (13%)	0
Single trial	1/6 (17%)	3/8 (38%)	1/3 (33%)
Analysis method in decision model			
Cost comparison	1/6 (17%)	0	0
Decision tree followed by Markov model	1/6 (17%)	2/8 (25%)	0
Discrete event simulation	0	0	1/3 (33%)
Markov model	2/6 (33%)	3/8 (38%)	0
Multi-state model	1/6 (17%)	2/8 (25%)	0
Partitioned survival	1/6 (17%)	0	0
UK Health Forum microsimulation	0	0	1/3 (33%)
Cost-utility analysis (model unspecified)	0/6	1/8 (13%)	1/3 (33%)
Software for cost-effectiveness			
Excel	2 (33%)	3 (38%)	1 (33%)
Not reported	3 (50%)	5 (63%)	0
R	0	0	1 (33%)

18

SAS	0	0	1 (33%)
Visual Basic & Excel	1 (17%)	0	0

Table 2: Results for time-to-event outcomes. NMA = network meta-analysis, PH = proportional hazards.

	NICE Technology Appraisal	NICE Guidelines	NIHR HTA Report
No. of articles	**19**	**4**	**3**
Clinical field			
Oncology	18 (95%)	2 (50%)	3 (100%)
Cardiovascular	0	1 (25%)	0
Kidney	0	1 (25%)	0
Multiple sclerosis	1 (5%)	0	0
Type of intervention			
Pharmacological	19 (100%)	2 (50%)	2 (67%)
Non-pharmacological	0	1 (25%)	1 (33%)
Both	0	1 (25%)	0
SYNTHESIS			
Method of synthesis			
Meta-analysis	4 (21%)	4 (100%)	1 (33%)
Network meta-analysis	16 (84%)	0	3 (100%)
Outcome measure			
Fractional polynomial	1 (5%)	0	0
Hazard ratio & fractional polynomial	1 (5%)	0	0
Hazard ratio	11 (58%)	4 (100%)	2 (67%)
Hazard ratio (PH & time-varying)	1 (5%)	0	0
Mean survival	1 (5%)	0	0
Median survival	2 (11%)	0	0
Median survival & hazard ratio	1 (5%)	0	0
Relative risk	0	0	1 (33%)
Time varying hazard ratio	1 (5%)	0	0
Fixed effect or random effects			
Fixed effect	2 (11%)	0	0
Random effects	0	0	1 (33%)
Both	8 (42%)	4 (100%)	1 (33%)
Unclear	9 (47%)	0	1 (33%)
Analysis details			
Multiple outcomes simultaneously	0	0	0
Lumping of interventions	1 (5%)	0	0
Multiple pairwise comparisons	3 (16%)	4 (100%)	1 (33%)
Multiple time points	1 (5%)	2 (50%)	0
Standardisation of any outcomes	1 (5%)	1 (25%)	0
Presentation of results			
Forest plot	7 (37%)	4 (100%)	2 (67%)
Network diagram (if NMA only)	11/16 (69%)	0/0	3/3 (100%)
Tables	13 (68%)	0	3 (100%)

20

Kaplan-Meier	10 (53%)	0	0
Other	3 (16%)	0	0
Software for synthesis			
Not reported	14 (74%)	0	1 (33%)
OpenBUGS	1 (5%)	0	0
R	1 (5%)	0	0
R & WinBUGS	1 (5%)	0	0
Cochrane Review Manager	0	4 (100%)	0
Stata	1 (5%)	0	0
Stata, R & JAGS	0	0	1 (33%)
WinBUGS	1 (5%)	0	1 (33%)
DECISION MODEL			
Decision model details			
Decision model included	19 (100%)	4 (100%)	3 (100%)
Graph of model structure	19 (100%)	4 (100%)	3 (100%)
Clinical results inform decision model	19 (100%)	4 (100%)	3 (100%)
How much clinical evidence informs decision model?			
Historical control	1 (5%)	0	0
Meta-analysis	2 (11%)	2 (50%)	1 (33%)
Network meta-analysis	12 (63%)	0	0
Single trial	4 (21%)	2 (50%)	2 (67%)
Analysis method in decision model			
Decision tree followed by Markov model	2 (11%)	1 (25%)	0
Markov model	1 (5%)	1 (25%)	0
Multi-state model	1 (5%)	0	0
Partitioned survival	14 (74%)	2 (50%)	2 (67%)
Semi-Markov cohort	1 (5%)	0	0
Weibull	0	0	1 (33%)
Software for cost-effectiveness			
Excel	9 (47%)	2 (50%)	3 (100%)
Not reported	10 (53%)	2 (50%)	0

21

```
┌─────────────────────────┐   ┌─────────────────────────┐   ┌─────────────────────────┐
│ 56 NICE technology      │   │ 27 NICE guidelines      │   │ 69 NIHR HTA reports     │
│ appraisals published    │   │ published April         │   │ published April         │
│ April 2018-March 2019   │   │ 2018-March 2019         │   │ 2018-March 2019         │
└─────────────────────────┘   └─────────────────────────┘   └─────────────────────────┘
```

18 No evidence synthesis 6 Terminated reviews 3 Update of previous review without evidence synthesis 2 Can't find company submission document 2 No continuous or time-to-event outcomes reported	9 No methods section for evidence synthesis 2 No continuous or time-to-event outcomes reported 1 No evidence synthesis	43 No evidence synthesis 7 Diagnostic 4 Narrative synthesis 3 Feasibility study 3 Prognostic 2 No continuous or time-to-event outcomes reported

47 reviews included (25 NICE technology appraisals, 15 NICE guidelines, 7 NIHR HTA reports)

Continuous outcome only	Time-to-event outcome only	Continuous & Time-to-event outcomes
6 NICE technology appraisals 11 NICE guidelines 4 NIHR HTA reports	18 NICE technology appraisals 2 NICE guidelines 3 NIHR HTA reports	1 NICE technology appraisal 2 NICE guidelines 0 NIHR HTA reports

What is new?

- Through a review of NIHR Health Technology Assessment reports, NICE technology appraisals and NICE guidelines published between 1st April 2018 and 31st March 2019 we establish that continuous and time-to-event outcomes are routinely synthesised within HTA articles. However, only two-thirds of articles used the results from an evidence synthesis of continuous or time-to-event outcomes to inform an economic decision model.

- Evidence from the review combined with expert knowledge of the methodological field identified five key areas where under-utilised methodology is available for improving the synthesis of continuous and time-to-event outcomes: i) outcomes reported on multiple scales, ii) reporting of multiple related outcomes, iii) appropriateness of the additive scale, iv) reporting of multiple time points and v) non-proportional hazards.

- We identified three anticipated barriers to the uptake and implementation of these methods: i) availability of specialist statistical expertise for model selection, model fitting and interpretation of results, ii) user-friendly software for implementing complex statistical and/or non-standard models, and iii) limited reporting of individual trials.

- More research is needed to develop, refine and generalise where possible methods for synthesising continuous and time-to-event outcomes for the purpose of decision modelling to maximise the evidence base utilised in the decision making process.

Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors.