ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Comparative-Effectiveness Research/HTA

A Guide to Selecting Flexible Survival Models to Inform Economic Evaluations of Cancer Immunotherapies

Stephen Palmer, MSc, Isabelle Borget, PhD, Tim Friede, PhD, Don Husereau, MSc, Jonathan Karnon, PhD, Ben Kearns, PhD, Emma Medin, MD, Elisabeth F.P. Peterse, PhD, Sven L. Klijn, MSc, Elisabeth J.M. Verburg-Baltussen, PhD, Elisabeth Fenwick, PhD, John Borrill, MSc

ABSTRACT

Objectives: Parametric models are routinely used to estimate the benefit of cancer drugs beyond trial follow-up. The advent of immune checkpoint inhibitors has challenged this paradigm, and emerging evidence suggests that more flexible survival models, which can better capture the shapes of complex hazard functions, might be needed for these interventions. Nevertheless, there is a need for an algorithm to help analysts decide whether flexible models are required and, if so, which should be chosen for testing. This position article has been produced to bridge this gap.

Methods: A virtual advisory board comprising 7 international experts with in-depth knowledge of survival analysis and health technology assessment was held in summer 2021. The experts discussed 24 questions across 6 topics: the current survival model selection procedure, data maturity, heterogeneity of treatment effect, cure and mortality, external evidence, and additions to existing guidelines. Their responses culminated in an algorithm to inform selection of flexible survival models.

Results: The algorithm consists of 8 steps and 4 questions. Key elements include the systematic identification of relevant external data, using clinical expert input at multiple points in the selection process, considering the future and the observed hazard functions, assessing the potential for long-term survivorship, and presenting results from all plausible models.

Conclusions: This algorithm provides a systematic, evidence-based approach to justify the selection of survival extrapolation models for cancer immunotherapies. If followed, it should reduce the risk of selecting inappropriate models, partially addressing a key area of uncertainty in the economic evaluation of these agents.

Keywords: algorithm, cancer, extrapolation, immunotherapy, survival analysis.

VALUE HEALTH. 2023; 26(2):185–192

Introduction

Parametric models are often used for extrapolating the long-term effects of cancer drugs across the entire time horizon of a cost-effectiveness analysis.[1-3] The family of standard parametric models that are commonly fitted to the Kaplan-Meier estimates of the survival function comprises exponential, Weibull, Gompertz, log-logistic, log-normal, and generalized gamma distributions. Each of these distributions specifies a particular shape for the hazard function. The exponential distribution assumes a constant hazard, whereas the Weibull and Gompertz distributions can reflect monotonically increasing or decreasing hazards, and the log-normal and log-logistic distributions are unimodal, whereas the generalized gamma distribution can assume a variety of shapes (eg, unimodal, monotonically increasing or decreasing, or bathtub).

Inappropriate parametric model selection can lead to unreliable estimates of incremental quality-adjusted life-years and, consequently, biased estimates of a treatment's cost-effectiveness. Therefore, analysts are required to rigorously justify their choice of extrapolation model. There is current guidance on how this should be done in a systematic and transparent way, including the National Institute for Health and Care Excellence Decision Support Unit technical support document 14.[4]

Treatment of advanced cancer with immunotherapy can produce deep and durable responses in a latent subgroup of patients. With sufficient trial follow-up, the empirical estimate of the mortality hazard rate (hereafter, the "observed hazard") with these agents is often found to change over time. The cancer-specific mortality risk is likely to increase at the start of treatment and then decline gradually in the medium term. In the longer term, the hazard may increase again, due to age-related mortality, if there is a nontrivial proportion of long-term survivors.[5] There is evidence to suggest that flexible survival models may provide an improved representation of this observed, and unobserved, survival function over that provided by standard parametric models.[6-8]

National Institute for Health and Care Excellence Decision Support Unit technical support document 21,[9] a guide on flexible methods for survival analysis, was recently published to address these issues and provide further support on trial-based approaches for extrapolation when time-varying hazards are encountered. It recommends extrapolating the treatment and control arms separately. It also encourages analysts to explicitly plot the assumed treatment effects in the short and long-term and the assumed hazards in the long term. Although detailed descriptions and limitations of a variety of flexible models are given, an algorithm to guide analysts on when to use these models and which to select for testing was not presented. This position article has been produced to address this gap.

Methods

A virtual advisory board comprising 7 international experts was held between June 23, 2021 and July 12, 2021. The experts were identified using a variety of approaches. Desk research was undertaken to identify individuals with expertise in survival analysis or health technology assessment (HTA) methodology. Publication histories were reviewed, as was any online information provided on their academic or professional profiles. Individuals were sought across a wide geographic area to capture a range of perspectives and opinions.

The virtual advisory panel was conducted using the Within3 platform. Topic selection was informed by an article by Quinn et al[10] on the challenges for assessing the long-term clinical benefit of cancer immunotherapy. Questions and relevant background materials were uploaded to the platform (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.07.009). Responses were visible to all participants. The experts had the opportunity to build on each other's ideas or offer alternative opinions. Moderators (E.F., S.L.K., and J.B.) could ask follow-up questions or seek clarification when needed. The platform was accessible by the experts and moderators at any time during the 3-week period the platform was open; this addressed the issue of experts being in different time zones. Individual transcripts of responses submitted by the 7 experts were downloaded from Within3 after the platform was closed, and consolidated into a single technical report that was shared with the experts for review. This article was developed from the content of this technical report. Advice was sought on extrapolation methods that could be applied to patient-level data from an unspecified immunotherapy study for an advanced or metastatic cancer.

Advice was sought on extrapolation methods that could be applied to patient-level data from an unspecified immunotherapy study for an advanced or metastatic cancer. Out-of-scope topics included surrogate endpoints for overall survival, evidence synthesis methods for time-varying hazards, and economic model frameworks.

The experts were asked to respond to and discuss 24 questions that had been divided into the following 6 topics: (1) the current selection procedure, (2) data maturity, (3) heterogeneity of treatment effect, (4) cure and mortality, (5) external evidence (ie, relevant survival data sourced outside an unspecified cancer immunotherapy clinical trial), and (6) additions to the existing guidelines for selecting appropriate models for survival extrapolation. The questions captured common challenges and reflected experiences of some of the authors in undertaking survival analyses for cancer immunotherapies. A complete overview of the questions presented to the experts is provided in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.07.009.

Based on the responses from the experts, an algorithm was developed for the selection of extrapolation models, with a focus on the use of flexible methods. This algorithm was shared with, and revised by, the experts in an iterative process until consensus was reached.

Results

The algorithm that was developed consists of 8 steps and 4 questions (see Fig. 1[9,11]). The key features of the model selection process are described in the sections below.

Importance of External Evidence in Extrapolation Model Selection (See Steps 1, 4, and 7)

It was agreed unanimously that the first step in the selection process should involve a targeted review to identify external evidence on the intervention and comparators. This evidence should be identified in a systematic and reproducible way. It is advised that particular attention should be paid to the patient characteristics, the length of follow-up, sample size, and the context and quality of the data to assess its relevance. A variety of tools are available to aid the systematic identification and assessment of the external evidence.[12,13]

Several categories of external evidence were identified. The first and most relevant source to consider is long-term survival data of the same products used in the same indication (eg, phase 1/2 trial data for the invention and/or randomized controlled trials, observation studies, or registry data for the comparators). The second option is to consider more mature data from the same products, but used in a later line of treatment for the same disease. The third is to assess evidence from a product with a similar mechanism of action used in the same indication. For example, analysts working on a new chimeric antigen receptor T-cell therapy may find long-term survival data reported for other chimeric antigen receptor T-cell therapies informative. Finally, insights may be gained from reviewing survival data of the same product used to treat other advanced cancers.

Several approaches have been proposed to formally use information from different sources for survival extrapolation modeling.[14-19] Nevertheless, these methods have not been standardized and the most appropriate method to use in any situation remains an area of ongoing research.[18]

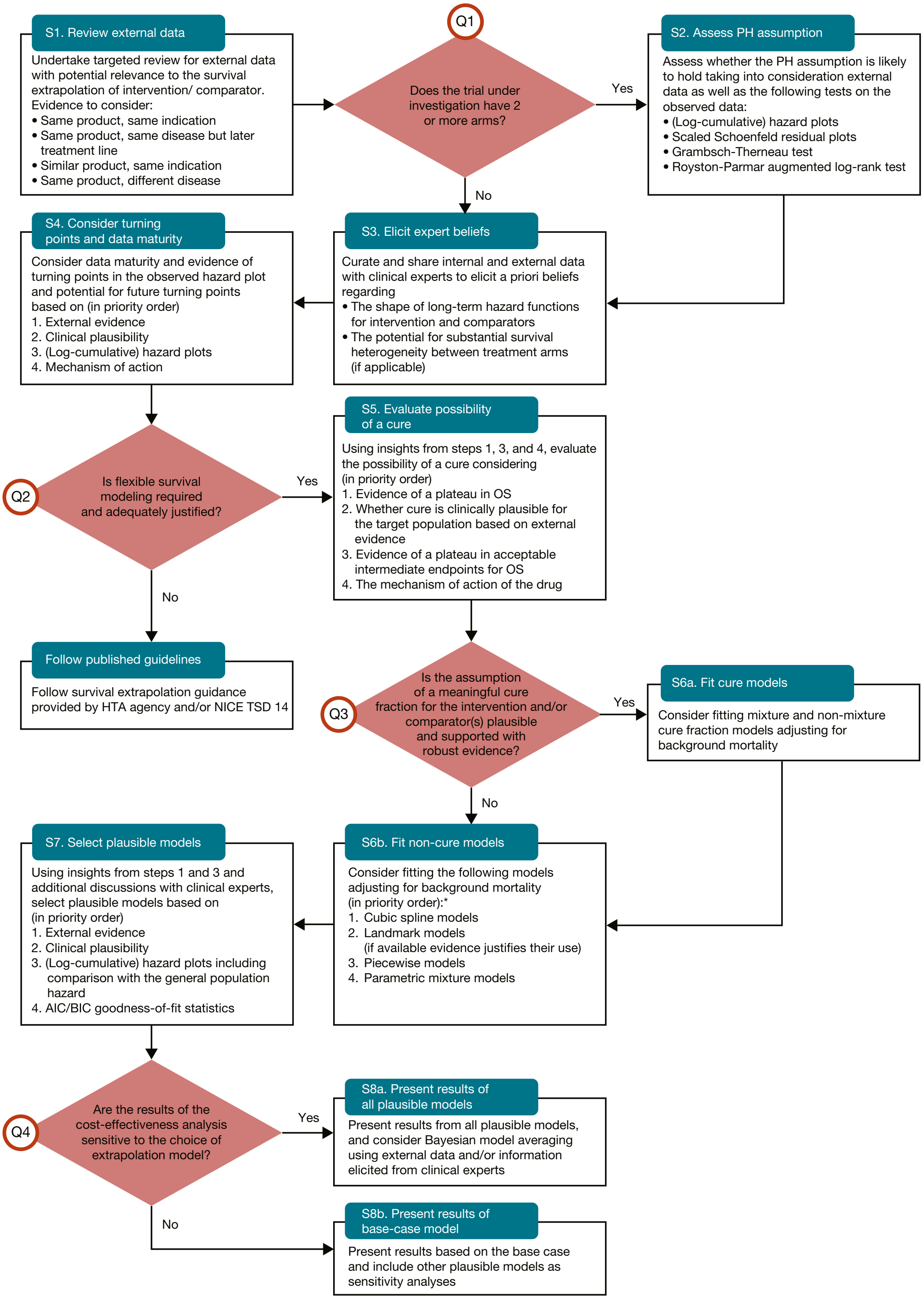
Use of Expert Opinion to Aid Model Selection (See Steps 3 and 7)

Clinical expert opinion should be sought before any model fitting to elicit beliefs on the likely shape of the hazard functions and expected survival in the medium and long terms (step 3). Eliciting expert opinion is also of value in helping to select and validate the most plausible models for the base-case analysis once those selected for testing have been fitted to the observed data (step 7).

It was acknowledged that there will be challenges in eliciting unbiased and meaningful judgments from clinical experts. It was also highlighted that there are no standard elicitation methods used to capture uncertainty in the experts' beliefs. The Sheffield Elicitation Framework[20] was one method the experts cited. This approach involves asking clinical experts to estimate mean, lower, and upper limits of landmark survival for a relevant patient population at, for example, 5, 10, and 20 years after the start of treatment.[8] In addition, analysts could refer to a recently published protocol of structured expert elicitation for healthcare decision making for further guidance.[21]

When fitting different extrapolation models to treatment arms, careful thought and clear justification should be given to address

Figure 1. Flexible survival model selection algorithm. Where there is a specific preference ordering, this is shown in the algorithm as numbered bullets with the accompanying text "in priority order." Note that all of the models described here could be implemented in a relative survival framework to take account of background mortality (ie, other-cause mortality).[9,11] *In addition, standard parametric models should also be considered for comparative purposes as HTA agencies are likely to expect to see them. Among the standard parametric models, the generalized gamma, log-logistic, and log-normal would be most suitable where flexible modeling is suggested given that the other models are not able to capture turning points in the hazard.


S1. Review external data
Undertake targeted review for external data with potential relevance to the survival extrapolation of intervention/ comparator. Evidence to consider:
• Same product, same indication
• Same product, same disease but later treatment line
• Similar product, same indication
• Same product, different disease
Q1
Does the trial under investigation have 2 or more arms?
Yes
No
S2. Assess PH assumption
Assess whether the PH assumption is likely to hold taking into consideration external data as well as the following tests on the observed data:
• (Log-cumulative) hazard plots
• Scaled Schoenfeld residual plots
• Grambsch-Therneau test
• Royston-Parmar augmented log-rank test
S4. Consider turning points and data maturity
Consider data maturity and evidence of turning points in the observed hazard plot and potential for future turning points based on (in priority order)
1. External evidence
2. Clinical plausibility
3. (Log-cumulative) hazard plots
4. Mechanism of action
S3. Elicit expert beliefs
Curate and share internal and external data with clinical experts to elicit a priori beliefs regarding
• The shape of long-term hazard functions for intervention and comparators
• The potential for substantial survival heterogeneity between treatment arms (if applicable)
Q2
Is flexible survival modeling required and adequately justified?
Yes
No
S5. Evaluate possibility of a cure
Using insights from steps 1, 3, and 4, evaluate the possibility of a cure considering (in priority order)
1. Evidence of a plateau in OS
2. Whether cure is clinically plausible for the target population based on external evidence
3. Evidence of a plateau in acceptable intermediate endpoints for OS
4. The mechanism of action of the drug
Follow published guidelines
Follow survival extrapolation guidance provided by HTA agency and/or NICE TSD 14
Q3
Is the assumption of a meaningful cure fraction for the intervention and/or comparator(s) plausible and supported with robust evidence?
Yes
No
S6a. Fit cure models
Consider fitting mixture and non-mixture cure fraction models adjusting for background mortality
S7. Select plausible models
Using insights from steps 1 and 3 and additional discussions with clinical experts, select plausible models based on (in priority order)
1. External evidence
2. Clinical plausibility
3. (Log-cumulative) hazard plots including comparison with the general population hazard
4. AIC/BIC goodness-of-fit statistics
S6b. Fit non-cure models
Consider fitting the following models adjusting for background mortality (in priority order):*
1. Cubic spline models
2. Landmark models (if available evidence justifies their use)
3. Piecewise models
4. Parametric mixture models
Q4
Are the results of the cost-effectiveness analysis sensitive to the choice of extrapolation model?
Yes
No
S8a. Present results of all plausible models
Present results from all plausible models, and consider Bayesian model averaging using external data and/or information elicited from clinical experts
S8b. Present results of base-case model
Present results based on the base case and include other plausible models as sensitivity analyses

HTA indicates health technology assessment; OS, overall survival; PHs, proportional hazards; Q, question; S, section.

concerns that survival estimates are not simply an artifact of the different models used. The rationale given could be informed by the external evidence identified in step 1, differences in mechanism of action of the treatment arms, and clinical expert judgment.

Analyses of Comparative Trial Data (See Steps 2, 4, and 5)

Analyses of comparative trial data are used for assessing the proportional hazards (PHs) assumption (step 2), to determine when flexible survival modeling is justified (step 4), and to evaluate the possibility of a cure (step 5).

Currently, the PHs assumption is assessed routinely to evaluate whether, in trials with 2 or more arms, a dependent model can be fitted (with a hazard ratio ideally applied to disease-specific mortality[22]) or whether independent models should be fitted to each arm.[4] The assessment of PHs should involve a combination of visual inspection of log-cumulative hazard plots and Schoenfeld residual plots (see Appendix Fig. 1 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.009) and statistical tests. The experts highlighted that the Grambsch-Therneau test[23] might be underpowered to detect violations of the PHs assumption. The Royston-Parmar augmented log-rank test was cited as an additional statistical test that could be performed.[24] It was noted that visual inspection of plots is subjective and no clear thresholds are available for determining, for example, whether the log-cumulative hazard plot can be considered parallel. Research is ongoing into incorporating time-varying treatment effects that relax the assumption of a constant treatment effect, and would remove the need to assess PHs.[25,26]

The potential for turning points in the hazard (ie, changes in the direction of the hazard function) should be evaluated to determine whether there is support for flexible survival modeling (step 4). It was highlighted that visual inspection to compare the (log-cumulative) hazard plots and assess the existence of turning points (either in the trial data or the external evidence) can be misleading. The application of different types of smoothing and changes to the kernel density used can have a significant impact on the smoothed hazards (Fig. 2). In particular, turning points observed at the tail of the survival data should be assessed for validity.[27] Therefore, both smoothed and unsmoothed hazard estimates should be inspected.[28]

In addition to the potential for turning points in the hazards, the maturity of the observed survival should be used to guide the decision of whether to consider flexible models. For illustration, an immunotherapy case study in advanced renal cell carcinoma has reported that a minimum follow-up of 39 months was needed for parametric mixture models and landmark models to provide reliable estimates of longer-term survival. Recent studies by Kearns et al,[29] Grant et al,[30] and Othus et al[31] have shown that the performance of cure modeling methods is heavily dependent on the maturity of the data. When data are immature, the use of landmark models, parametric mixture models, and cure models may only be warranted if robust external data are available to inform the models.

Evidence of a sustained plateau in overall survival may be indicative of statistical cure (step 5). Nevertheless, if data are heavily right censored and the sample size is small, this visual inspection can be misleading. For treatments suspected to have potential curative properties, justification for using a cure model is likely to require external sources of data given that, at time of HTA submission, trials typically have a median follow-up of $<$ 24 months. This is considerably less than the minimum length of follow-up required to reliably estimate statistical cure for many cancers.[32] When validated predictive markers for overall survival are available, which can be assessed in the relatively short timeframe, this could help support claims of statistical cure.[33]

Selection of Plausible Models (See Steps 6a, 6b, and 7)

Mixture and nonmixture cure fraction models could be considered when a meaningful cure fraction is plausible and supported by robust evidence (step 6a). Nevertheless, the use of cure models for HTA decision making is an area of contention, and there is no clear definition of when a cure fraction might be considered "meaningful." In discussions, one expert suggested using 5% as a ballpark estimate, provided that the sample size and duration of follow-up are adequate. To estimate the cure fraction reliably, it is essential that there are sufficient numbers at risk in the plateau phase of the overall survival/intermediate endpoint curves.

Irrespective of the potential for cure, the following models (step 6b) should be considered: (1) cubic spline models; (2) landmark models, if available evidence justifies their use; (3) piecewise models; and (4) parametric mixture models. In addition, standard parametric models should also be fitted. This was considered necessary in the interest of model parsimony. HTA agencies would want to see that a complex model had not been selected when a simpler one could have been equally justified. Important considerations for each of these model types are provided in Table 1.[9,34]

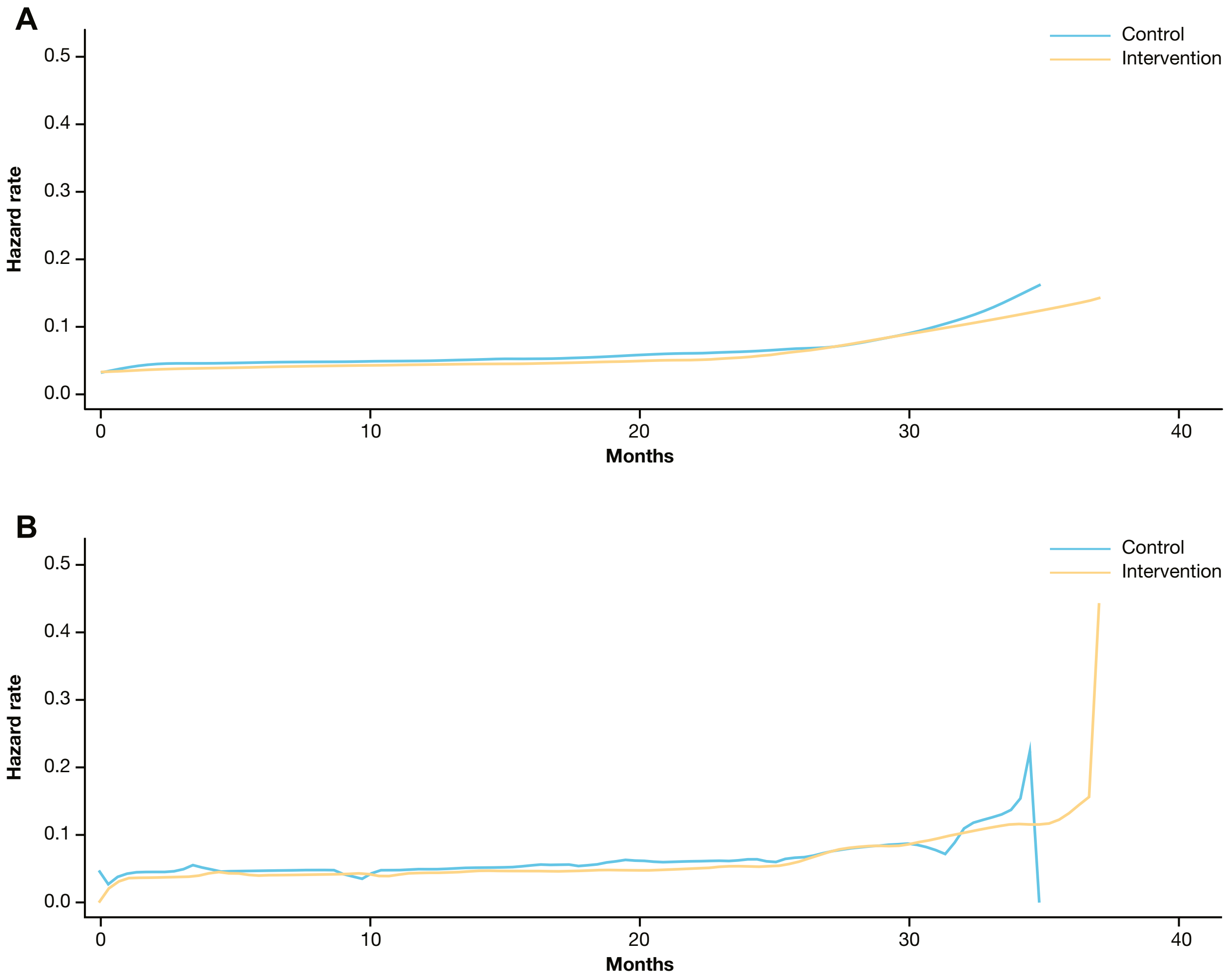
Visual comparison of hazard estimates, used to select plausible models (step 7), should be interpreted with caution. Often when inspecting hazard plots and Kaplan-Meier estimates of the survival function, individuals are inclined to give equal weight to all time points, regardless of the number of patients at risk. Analysts may be tempted to consider that a flexible model is required because standard models appear to poorly fit the tails of these plots. Presenting the number of patients at risk at relevant time points and the confidence intervals is recommended to prevent this from happening. In addition, a suggestion was made to restrict figures of the plot of the hazards inherent to the trial data (observed hazard) and Kaplan-Meier estimators to a point in time when the numbers at risk are still "reasonable." If this practice were to be followed, it was agreed the time point should be justified, and these time-restricted plots presented alongside those with the complete data. Limited guidance is offered to help inform the time point at which the plots are curtailed. Pocock et al[35] suggested a point at which 10% to 20% of the patients are still at risk. More recently, Gebski et al[36] suggested a less conservative approach, using the size of the decrease of the percent survival estimate at time point t if 1 extra event should occur immediately after t.

The experts believed that information criteria (eg, Akaike's information criterion [AIC]/Bayesian information criterion [BIC] goodness-of-fit test statistics) should have the lowest priority for selecting plausible models, although they might still be useful for exclusion purposes. All experts rated the importance of the AIC/BIC statistic below 5, on a scale of 1 (not important) to 10 (extremely important), with an average score, across all experts, of 3.7. Their opinion may be at odds with current practice. Nevertheless, as noted by several experts, "a good statistical fit to the observed data does not necessarily ensure a model will provide an accurate estimate of long-term survival." In a recent simulation study, Kearns et al[33] reported that models with excellent within-sample fit could sometimes provide poor extrapolation performance.

Presentation of Results (See Steps 8a and 8b)

When several extrapolation models have been identified that are all clinically plausible, cost-effectiveness results should be

Figure 2. The potential impact associated with alternative smoothing algorithms. (A) Smoothed hazard. (B) Smoothed hazards with different smoothing. The potential impact of using a different smoothing algorithm with a different kernel density. This is based on generated data and not related to any clinical trial or real-world data. In panel A (bandwidth = 5), hazard rates for both treatment arms are fairly stable for the initial 24 months with both increasing slightly from ∼25 months onward. Nevertheless, in panel B (bandwidth = 2), although the hazards are again stable for the initial period followed by an increase at approximately 25 months, there is a sharp downward spike for the control arm at ∼35 months and a sharp upward spike for the intervention arm at ∼38 months. These spikes are artifacts of the different smoothing window (kernel density) used.



presented to decision makers for all of these models to show the impact of structural uncertainty (step 8a).

Bayesian model averaging was proposed as an approach that could be considered when multiple models could be selected for the base-case analysis.[37] Nevertheless, current HTA acceptance of this method is uncertain, and determining the weights that should be applied in the Bayesian model averaging approach, as well as how to incorporate them in a probabilistic analysis, remains an area of ongoing research.[38]

In the situation when the cost-effectiveness results are insensitive to the choice of extrapolation model, the most plausible model should be selected to represent the base case (step 8b). All other plausible alternatives should then be presented as scenario analyses.

Irrespective of how the results are presented, in cases when different extrapolation methods were used for different treatment arms, analysts should be explicit about the rationale for the different extrapolation methods. The biological and clinical plausibility of the models should be discussed. It was also suggested that hazard ratios at different time points should be presented so that decision makers can better judge whether this modeling approach was reasonable.

Discussion

Analysts conducting economic evaluations of cancer immunotherapies now have a wide range of extrapolation models at their disposal to estimate long-term survival. Appropriate and fully justified model selection is crucial to ensure that reliable estimates of quality-adjusted life-years are used to inform cost-effectiveness analysis. An international panel of experts was convened to offer guidance on selection of flexible survival models, taking into consideration the current positions taken by HTA agencies and the most recent developments in this field.

Table 1. Overview of different types of models to consider and important considerations for each model.

Type of model	Considerations
Models to consider when a meaningful cure fraction is plausible and supported by robust evidence	
1. Mixture cure fraction models	Assumes there are 2 groups of individuals in a population: those cured of their disease, who follow general population mortality, and those uncured, who follow a disease-specific survival function[9]
2. Nonmixture cure fraction models	Does not assume the population is separated into groups, but rather a mathematical function is used to define an asymptote for the survival function as time tends to infinity (ie, the point at which the cause-specific mortality becomes zero)
Models to consider irrespective of the potential for a cure	
1. Cubic spline models	Of the flexible models, the cubic spline models require relatively weak assumptions because they do not rely on the assumption of certain subgroups. Given that placement of the knots can be highly subjective, it is advised to place the knots uniformly along the distribution of uncensored log event times, with boundary knots placed at the minimum and maximum.[9]
2. Landmark models	A strength of this model is that it has intuitive appeal. Nevertheless, selecting the landmark time point and the choice of response measure can be challenging and difficult to justify. Therefore, these models are only suitable when there is a well-established landmark time and a strong link between response and survival.
3. Piecewise models	The challenge with these models is to determine the cut points given that these can influence the results.[34] Furthermore, these models imply a sudden change in the hazard, which might be clinically impossible and lack intuitive appeal.
4. Parametric mixture models	These models explicitly acknowledge the heterogeneity in survival. Nevertheless, they require a relatively large number of parameters to be estimated, with the risk of identification problems and overfitting. Using parametric mixture models has not been routinely accepted by health technology assessment agencies.
Models to fit for comparative purposes	
1. Standard parametric models	The generalized gamma, log-logistic, and log-normal would be the most suitable where flexible modeling is suggested given that the other models are not able to capture turning points in the hazard. Standard parametric models typically avoid overfitting to the data but can sometimes fail to adequately describe observed hazard patterns.

External evidence and its use to inform appropriate flexible model selection was a key topic of debate. Multiple articles on this subject were cited.[14-19] When using external evidence, it is important to assess how closely the survival observed in the trial population is likely to mirror the long-term external data that have been identified. Comprehensive guidance on how to incorporate different types of external evidence is lacking and is an important area of future research. This includes how best to elicit information from clinical experts to inform survival extrapolations. The experts also encourage analysts and decision makers not to rely solely on AIC/BIC goodness-of-fit statistics to inform model selection. Careful consideration of the totality of the available evidence should be given when making decisions regarding extrapolation.

It is evident that there is no "one size fits all" modeling solution and that multiple plausible extrapolation models may need to be considered and examined. Nevertheless, it was also accepted that testing all conceivable models would be inefficient and time consuming. Analysts may elect to apply constraints to reduce the analytical burden and avoid overfitting, especially when using parametric mixture models. To aid the selection process, the existence of one or more turning points in the hazard function would preclude the use of extrapolation models that assume constant or monotonically increasing, or decreasing, hazards. Immature data, the lack of external data, and small trial sample size may also rule out the use of more flexible models such as mixture cure, parametric mixture, and response-based landmark models.[6,8,31] When there is evidence of time-varying hazards, all of the experts agreed that cubic spline models should be examined routinely. This recommendation is supported by a recent study that evaluated how well standard parametric and spline models predicted survival when fitted to cancer registry data with artificially right-censored follow-up times.[7] The researchers of this study found that, across all data sets, spline odds and spline normal models frequently gave more accurate predictions of 10-year survival than standard parametric models. It is also worthwhile noting that a Bayesian network meta-analysis approach exists that uses restricted cubic splines to address the problem of time-varying hazards.[39] This is an important consideration, given that estimates of relative treatment effects generated by this network meta-analysis could be readily incorporated into cost-effectiveness models.

Several limitations to this guide for flexible survival model selection have been identified. Because of the time constraints associated with a virtual advisory board, only 6 predefined topics were open for discussion. Several key topics related to extrapolation, such as the appropriate use of surrogate endpoints and the application of excess mortality to inform long-term survival projections, remain to be addressed. The authors are also aware that survival analysis is an active field of research. As a result, the presented algorithm would require periodic updates as emerging methods gain traction. Several promising areas of research have been identified for possible future inclusion. Joint survival modeling and dynamic survival models that use time-dependent coefficients (eg, response, treatment discontinuation, and progression) in a Cox regression model to predict future hazards are currently under evaluation.[40,41] The application of Bayesian modeling methods for combining evidence from multiple sources in extrapolation models builds on work published as early as 2006.[42] One approach uses a Bayesian method that allows more mature external data to first guide the parametric model selection and then to provide a priori distributions that are used to inform the shape parameter of the parametric distribution fitted to the pivotal trial data.[16]

Conclusion

The algorithm presented in this article provides a systematic and evidence-based approach to justify the selection of survival models, while also considering the latest developments in survival analysis. This algorithm may also be used to aid in critiques of approaches to survival model selection. If followed, it should ensure the use of appropriate models, identifying when more complex approaches are required and avoiding the use of overly sophisticated techniques when standard models would suffice. We believe the algorithm will also improve transparency and consistency, leading to increased confidence in economic evaluations of cancer immunotherapies. It may also prove useful for other treatments and diseases where more flexible extrapolation models may be warranted.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.07.009.

Article and Author Information

Accepted for Publication: July 6, 2022

Published Online: August 13, 2022

doi: https://doi.org/10.1016/j.jval.2022.07.009

Author Affiliations: Centre for Health Economics, University of York, York, England, UK (Palmer); Biostatistics and Epidemiology Office, Gustave Roussy, Paris-Saclay University, Villejuif, France (Borget); Oncostat, Paris-Saclay University U1018, Inserm, Paris-Saclay University, “Ligue Contre le Cancer” labeled team, Villejuif, France (Borget); Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany (Friede); School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada (Husereau); Flinders Health and Medical Research Institute, Flinders University, Adelaide, SA, Australia (Karnon); School of Health and Related Research, University of Sheffield, Sheffield, England, UK (Kearns); Parexel International, Stockholm, Sweden (Medin); Department of Learning, Infomatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden (Medin); Modeling & Meta-analysis, OPEN Health, Rotterdam, The Netherlands (Peterse, Verburg-Baltussen); Worldwide Health Economics and Outcomes Research - Economic and Predictive Modeling, Bristol Myers Squibb, Utrecht, The Netherlands (Klijn); Modeling & Meta-analysis, OPEN Health, Oxford, England, UK (Fenwick); Worldwide Health Economics and Outcomes Research, Bristol Myers Squibb, Uxbridge, Greater London, England, UK (Borrill).

Correspondence: John Borrill, MSc, Bristol Myers Squibb, Uxbridge Business Park, Sanderson Road, Uxbridge, England UB8 1DH, United Kingdom. Email: john.borrill@bms.com

Author Contributions: *Concept and design:* Medin, Klijn, Verburg-Baltussen, Borrill
Acquisition of data: Palmer, Borget, Friede, Husereau, Karnon, Kearns, Medin
Analysis and interpretation of data: Palmer, Borget, Friede, Husereau, Karnon, Kearns, Medin, Peterse, Klijn, Verburg-Baltussen, Fenwick, Borrill
Drafting of the manuscript: Palmer, Borget, Friede, Husereau, Karnon, Kearns, Medin, Peterse, Klijn, Verburg-Baltussen, Fenwick, Borrill
Critical revision of the paper for important intellectual content: Palmer, Borget, Friede, Husereau, Karnon, Kearns, Medin, Peterse, Klijn, Fenwick, Borrill
Obtaining funding: Borrill
Administrative, technical, or logistic support: Verburg-Baltussen, Fenwick, Borrill
Supervision: Klijn, Fenwick, Borrill

Conflict of Interest Disclosures: Messrs Palmer, Husereau, and Borrill, and Drs Borget, Friede, Karnon, Kearns, Medin, Peterse, and Verburg-Baltussen reported receiving personal fees from Bristol Myers Squibb during the conduct of the study. Dr Friede reported receiving personal fees from Bayer, CSL Behring, Galapagos, Minoryx, Novartis, Roche, Vifor, LivaNova, Fresenius Kabi, BiosenseWebster, Janssen, Enanta, Eli Lilly, and IQVIA outside the submitted work. Mr Husereau reported receiving personal fees from GSK and Roche outside the submitted work. Dr Kearns reported receiving grants from National Institute for Health and Care Research outside the submitted work. Mr Klijn reported receiving grants and other fees from Bristol Myers Squibb during the conduct of the study. Dr Fenwick reported receiving grants from Bristol Myers Squibb during the conduct of the study. No other disclosures were reported.

Funding/Support: This research was supported by Bristol Myers Squibb.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the project; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgment: The authors would like to thank Lisette Nientker for reviewing the meeting transcripts and Lucie Gower and Annika Voss for technical support with the Within3 platform. Editorial support was provided by Parexel.

REFERENCES

1. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (version 5.0). Australian Government Department of Health. https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf. Accessed August 9, 2022.
2. Choices in methods for economic evaluations. Haute Autorité de Santé. https://www.has-sante.fr/jcms/r_1499251/en/choices-in-methods-for-economic-evaluation. Accessed August 9, 2022.
3. Guidelines for the economic evaluation of health technologies: Canada. Canadian Agency for Drugs and Technologies in Health. https://www.cadth.ca/about-cadth/how-we-do-it/methods-and-guidelines/guidelines-for-the-economic-evaluation-of-health-technologies-Canada. Accessed August 9, 2022.
4. Latimer N. NICE DSU technical support document 14: undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. http://www.nicedsu.org.uk. Accessed August 9, 2022.
5. Ouwens MJNM, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. *Pharmacoeconomics*. 2019;37(9):1129–1138.
6. Bullement A, Willis A, Amin A, Schlichting M, Hatswell AJ, Bharmal M. Evaluation of survival extrapolation in immuno-oncology using multiple preplanned data cuts: learnings to aid in model selection. *BMC Med Res Methodol*. 2020;20(1):103.
7. Gray J, Sullivan T, Latimer NR, et al. Extrapolation of survival curves using standard parametric models and flexible parametric spline models: comparisons in large registry cohorts with advanced cancer. *Med Decis Making*. 2021;41(2):179–193.
8. Klijn SL, Fenwick E, Kroep S, et al. What did time tell us? A comparison and retrospective validation of different survival extrapolation methods for immuno-oncologic therapy in advanced or metastatic renal cell carcinoma. *Pharmacoeconomics*. 2021;39(3):345–356.
9. Rutherford MJ, Lambert PC, Sweeting MJ, et al. NICE DSU technical support document 21. Flexible methods for survival analysis. http://www.nicedsu.org.uk. Accessed August 9, 2022.
10. Quinn C, Garrison LP, Pownell AK, et al. Current challenges for assessing the long-term clinical benefit of cancer immunotherapy: a multi-stakeholder perspective. *J Immunother Cancer*. 2020;8(2):e000648.
11. Andersson TM, Dickman PW, Eloranta S, Lambert PC. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Med Res Methodol*. 2011;11:96.
12. Sterne JA, Hernan MA, Reeves BC, et al. Robins-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
13. Higgins JPT, Thomas J, Chandler J, et al. Cochrane handbook for systematic reviews of interventions (version 6.2). www.training.cochrane.org/handbook. Accessed August 9, 2022.
14. Boatman JA, Vock DM, Koopmeiners JS. Borrowing from supplemental sources to estimate causal effects from a primary data source. *Stat Med*. 2021;40(24):5115–5130.

15. Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of survival curves from cancer trials using external information. *Med Decis Making*. 2017;37(4):353–366.
16. Soikkeli F, Hashim M, Ouwens M, Postma M, Heeg B. Extrapolating survival data using historical trial-based a priori distributions. *Value Health*. 2019;22(9):1012–1017.
17. Papanikos T, Thompson JR, Abrams KR, et al. Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data. *Stat Med*. 2020;39(8):1103–1124.
18. Nikolaidis GF, Woods B, Palmer S, Soares MO. Classifying information-sharing methods. *BMC Med Res Methodol*. 2021;21(1):107.
19. Jackson C, Stevens J, Ren S, et al. Extrapolating survival from randomized trials using external data: a review of methods. *Med Decis Making*. 2017;37(4):377–390.
20. Gosling JP. SHELF: the Sheffield elicitation framework. In: Dias L, Morton A, Quigley J, eds. *Elicitation: The Science and Art of Structuring Judgement*. Cham, Switzerland: Springer; 2018:61–93.
21. Bojke L, Soares M, Claxton K, et al. Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technol Assess*. 2021;25(37):1–124.
22. Alarid-Escudero F, Kuntz KM. Potential bias associated with modeling the effectiveness of healthcare interventions in reducing mortality using an overall hazard ratio. *Pharmacoeconomics*. 2020;38(3):285–296.
23. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):512–526.
24. Royston P, Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Med Res Methodol*. 2016;16:16.
25. Jacobs IJ, Menon U, Ryan A, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial [published correction appears in *Lancet*. 2016;387(10022):944]. *Lancet*. 2016;387(10022):945–956.
26. Kearns B, Chilcott J, Whyte S, Preston L, Sadler S. Cost-effectiveness of screening for ovarian cancer amongst postmenopausal women: a model-based economic evaluation [published correction appears in *BMC Med*. 2017;15(1):31]. *BMC Med*. 2016;14(1):200.
27. Wang J-L. Smoothing hazard rates: contribution to the encyclopedia of biostatistics. https://anson.ucdavis.edu/~wang/hazardsmoothing.pdf. Accessed August 9, 2022.
28. Kearns B, Stevens J, Ren S, Brennan A. How uncertain is the survival extrapolation? A study of the impact of different parametric survival models on extrapolated uncertainty about hazard functions, lifetime mean survival and cost effectiveness. *Pharmacoeconomics*. 2020;38(2):193–204.
29. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. The extrapolation performance of survival models for data with a cure fraction: a simulation study. *Value Health*. 2021;24(11):1634–1642.
30. Grant TS, Burns D, Kiff C, Lee D. A case study examining the usefulness of cure modelling for the prediction of survival based on data maturity. *Pharmacoeconomics*. 2020;38(4):385–395.
31. Othus M, Bansal A, Erba H, Ramsey S. Bias in mean survival from fitting cure models with limited follow-up. *Value Health*. 2020;23(8):1034–1039.
32. Tai P, Yu E, Cserni G, et al. Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database. *BMC Cancer*. 2005;5:48.
33. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. Comparing current and emerging practice models for the extrapolation of survival data: a simulation study and case-study. *BMC Med Res Methodol*. 2021;21(1):263.
34. Davies A, Briggs A, Schneider J, et al. The ends justify the mean: outcome measures for estimating the value of new cancer therapies. *Health Outcomes Res Med*. 2012;3(1):e25–e36.
35. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet*. 2002;359(9318):1686–1689.
36. Gebski V, Gares V, Gibbs E, Byth K. Data maturity and follow-up in time-to-event analyses. *Int J Epidemiol*. 2018;47(3):850–859.
37. Jackson CH, Thompson SG, Sharples LD. Accounting for uncertainty in health economic decision models by using model averaging. *J R Stat Soc Ser A Stat Soc*. 2009;172(2):383–404.
38. Hardern C, Lee D, Sly I, Kearns B. Structural uncertainty in survival extrapolation: exploring the impact of four model averaging methods and adjusting for data maturity. *Value Health*. 2020;23(Suppl 2):S402.
39. 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.
40. Crowther MJ, Lambert PC, Abrams KR. Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: a joint modelling approach. *BMC Med Res Methodol*. 2013;13:146.
41. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. Generalized linear models for flexible parametric modeling of the hazard function. *Med Decis Making*. 2019;39(7):867–878.
42. Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Stat Med*. 2006;25(11):1960–1975.