# NICE DSU TECHNICAL SUPPORT DOCUMENT 21: Flexible Methods for Survival Analysis

REPORT BY THE DECISION SUPPORT UNIT

23 January 2020

Mark J Rutherford[1], Paul C Lambert[1,2], Michael J Sweeting[1], Becky Pennington[3], Michael J Crowther[1,2], Keith R Abrams[1], Nicholas R Latimer[3]

[1] Department of Health Sciences, University of Leicester, Leicester, UK

[2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[3] School of Health and Related Research, University of Sheffield, Sheffield, UK

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street
Sheffield, S1 4DA UK

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

## ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

### ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

The NICE Guide to the Methods of Technology Appraisal is a regularly updated document
that provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The Methods Guide does not provide detailed advice on how to implement and apply the methods it describes. This DSU series of Technical Support Documents (TSDs) is intended to complement the Methods Guide by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in each topic area, and make clear recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE Technology Appraisals, whether companies, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute

formal NICE guidance or policy.

Prof Allan Wailoo, Director of DSU and TSD series editor.

[i] National Institute for Health and Care Excellence. Guide to the methods of technology appraisal, 2013 (updated April 2013), London

**ACKNOWLEDGEMENTS:**

The authors would like to acknowledge the contributions of the independent reviewers; Adrian Bagust, Chris Jackson, John Stevens, and the reviewers from NICE; Nicola Hay, Christian Griffiths and Jamie Elvidge. The peer-review comments helped to improve the clarity and quality of this document.

**This report should be referenced as follows:**

Rutherford, MJ., Lambert, PC., Sweeting, MJ., Pennington, R., Crowther, MJ., Abrams, KR., Latimer, NR. NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis. 2020 [Available from http://www.nicedsu.org.uk]

# EXECUTIVE SUMMARY

Survival analysis modelling approaches are often required to capture the survival functions seen in clinical trial data and to further extrapolate to estimate lifetime benefits in economic evaluations. To accurately capture overall survival functions, there is a need to consider the hazards (rates of events) both within the trial period and beyond the duration of the trial. Often complex hazard function shapes can arise both within and beyond the trial period, meaning that increasingly sophisticated survival models are required and are being applied in NICE TAs, going beyond standard parametric survival models. In particular, the advent of immuno-therapy treatments for oncology has resulted in an increase in the use of complex survival models, because delayed responses to treatment and the existence of long-term survivors have been hypothesised to result in complex hazard functions.

We take a single trial-based approach to this issue in this document. That is, we focus on methods that could be applied to data within a trial and then used (with the potential inclusion of external data) to provide a measure of the mean survival for that specific trial population. If the target population differs from those included in the trial, then methods to account for the different make-up of the population need to be accounted for when estimating absolute effect measures. We further concentrate on extrapolating treatment and control arms separately, and thereby have a treatment effect that is implied by those two separate extrapolations rather than directly modelling the treatment effect (and whether this may diminish or stabilise). We argue that one should explicitly plot the assumed long-term hazard in each trial arm and to plot the assumed treatment effect in the short and long-term. In taking this single trial-based approach, we do not cover the specific requirements needed when pooling evidence across studies – extra care is required when estimating absolute rather than relative effects in this context. Many of the issues outlined above are covered in separate, specific TSD documents.

We describe a variety of survival modelling approaches that have been, and can be, used, when hazard functions are complex. Flexible parametric survival methods incorporating splines or fractional polynomials, models that enforce cure proportions, and more general mixture models have been applied in NICE TAs in the presence of

complexity of observed hazard functions. Further approaches that have been used in practice take a conditional approach to dealing with the issue of complex hazard functions, namely piecewise modelling approaches and landmarking on a point of treatment response. We present the motivation behind each approach, their details with respect to formulae and assumptions, and their limitations, all from the perspective of applying them to observed trial data. We further demonstrate their potential performance in a range of plausible and realistic simulated scenarios. However, a major consideration remains surrounding how to then extrapolate survival functions to a lifetime horizon, whether this is using standard parametric approaches or a more sophisticated survival modelling technique. Extensive discussion and consideration of the assumptions that are made under various approaches when extrapolating survival functions beyond the range of the data are therefore also given.

Determining whether a complex survival model provides a good representation of hazards beyond the trial period is challenging, and general issues with extrapolation of survival functions still apply. Careful consideration should therefore be given to whether the extrapolation is realistic, which may involve using external data sources, clinical expert opinion, or arguments around biological plausibility (incorporating knowledge around expected shapes of survival functions in given populations, or knowledge of specific disease subgroup characteristics). However, retrospectively assessing the plausibility of extrapolations is inherently subjective and as a result may be prone to personal bias. Formal prior elicitation of likely long-term survival functions, potentially using a Bayesian framework for the technology assessment, is one possible solution, though not one that we cover in detail in this document nor one which has been extensively evaluated or used in practice, and is not one that necessarily avoids the issue of subjectivity.

We apply each of the survival modelling approaches to complex simulated survival data; simulating from scenarios with turning points in the true hazard functions driven by competing risks of both disease-specific and other-cause mortality. We evaluate each of the described approaches in capturing the true survival functions both within the range of follow-up and extrapolated to a lifetime horizon. We discuss when, and why, both the simple and complex approaches fail, and further when the incorporation of external data may improve long-term extrapolation. We also provide illustrative examples of some of the issues highlighted in the simulation study and discuss implications for practice.

We provide specific recommendations for consideration for each of the considered complex approaches; both in their fitting to observed data and the consideration of how to evaluate extrapolated survival functions. These relate to model selection and convergence issues, the choice of cut-points should they be required, and when, and if, one should incorporate external data. We further provide general recommendations for all approaches; considering the incorporation of external data for extrapolation of survival functions, approaches for evaluating if model-based extrapolations are consistent with external data, and graphical approaches to allow the assessment of the appropriateness of the assumed treatment effect in the extrapolated portion of a survival function. We also provide recommendations for areas of further research and methods evaluation.

# 1. CONTENTS

1. Contents ........ 8
2. INTRODUCTION ........ 12
   2.1. Background ........ 12
      2.1.1. Survival analysis in economic evaluation ........ 12
      2.1.2. Hazard functions ........ 14
      2.1.3. Use of flexible models in NICE Technology Appraisals ........ 16
3. METHODS: FITTING TO OBSERVED DATA ........ 20
   3.1. Overview ........ 20
   3.2. Flexible parametric survival models ........ 20
      3.2.1. Motivation ........ 20
      3.2.2. Details ........ 20
      3.2.3. Limitations ........ 24
      3.2.4. Further Issues ........ 24
      3.2.5. Other flexible parametric models ........ 25
   3.3. Mixture models ........ 26
      3.3.1. Motivation ........ 26
      3.3.2. Details ........ 27
      3.3.3. Limitations ........ 28
      3.3.4. Further Issues ........ 29
   3.4. LANDMARK MODELS ........ 29
      3.4.1. Motivation ........ 29
      3.4.2. Details ........ 30
      3.4.3. Limitations ........ 31
      3.4.4. Further issues ........ 32
   3.5. PIECEWISE MODELS ........ 32
      3.5.1. Motivation ........ 32
      3.5.2. Details ........ 34
      3.5.3. Limitations ........ 35
      3.5.4. Further issues ........ 36
   3.6. CURE MODELS ........ 36
      3.6.1. Motivation ........ 36
      3.6.2. Details ........ 37
      3.6.3. Limitations ........ 40
      3.6.4. Further issues ........ 41
   3.7. EXCESS MORTALITY / RELATIVE SURVIVAL MODELS ........ 41

3.7.1. Motivation ....................................................41
3.7.2. Details ....................................................42
3.7.3. Limitations ....................................................43
3.7.4. Further issues ....................................................43
3.8. ADOPTING A BAYESIAN APPROACH ....................................................44
4. METHODS: EXTRAPOLATION OF SURVIVAL FUNCTIONS AND INCORPORATING EXTERNAL INFORMATION....................................................46
4.1. BACKGROUND ....................................................46
4.2. INCORPORATING EXTERNAL DATA ....................................................47
4.3. EXTRAPOLATION FOR COMPLEX SURVIVAL APPROACHES ....................................................49
4.3.1. Flexible parametric models....................................................49
4.3.2. Mixture models ....................................................50
4.3.3. Landmark....................................................50
4.3.4. Piecewise....................................................51
4.3.5. Cure....................................................51
4.3.6. Excess mortality / cause-specific mortality (relative survival) ....................................................51
4.4. CONSIDERATIONS FOR ALL EXTERNAL DATA APPROACHES ....................................................52
5. SIMULATION STUDIES ....................................................55
5.1. INTRODUCTION ....................................................55
5.2. METHODS....................................................55
5.2.1. Data generating mechanisms ....................................................55
5.2.2. Estimands....................................................58
5.2.3. Survival models ....................................................66
5.2.4. Performance measures....................................................67
5.3. RESULTS ....................................................67
5.3.1. Standard Models ....................................................71
5.3.2. Flexible Parametric Models ....................................................71
5.3.3. Flexible Parametric Models (incorporating background mortality) ....................................................72
5.3.4. Cure Models (incorporates background mortality)....................................................73
5.4. SUMMARY & DISCUSSION....................................................73
6. EXAMPLES OF MODEL EXTRAPOLATION FOR CHOSEN SIMULATION SCENARIOS ....................................................75
6.1. introduction ....................................................75
6.2. calculating marginal expected survival and hazard ....................................................75
6.3. Assessing model fit within the range of the data does not guarantee good extrapolation....................................................76
6.4. Incorporating background mortality into modelling may help with extrapolation....................................................79

6.5. What might happen if incorrect background mortality rates are used? .......... 81
6.6. Having longer-term follow-up information is likely to be beneficial, but not always for all-cause approach .......... 83
6.7. Altering the time to cure in order to use the population hazard only in the long-term .......... 85
6.8. Piecewise models; Reduction of sample size at later cut-points .......... 86
7. DISCUSSION AND RECOMMENDATIONS .......... 88
7.1. General Recommendations: .......... 88
7.2. Model-specific recommendations: .......... 90
7.3. Recommendations for Research: .......... 91
APPENDIX .......... 93
A.1 Stata code to generate simulation study data .......... 93
A.2 STATA CODE USED TO UNDERTAKE ANALYSES .......... 93
A.3 STATA CODE TO CALCULATE MARGINAL EXPECTED SURVIVAL AND HAZARD .......... 93
A.4 SIMULATION STUDY RESULTS .......... 93
REFERENCES .......... 94

## FIGURES

Figure 1: Hazard functions associated with Weibull and exponential models. The shape of the Weibull hazard function is determined by its shape parameter, $\boldsymbol{\gamma}$. .......... 13
Figure 2: Hazard functions associated with Gompertz models. The shape of the Gompertz hazard function is determined by its shape parameter, $\boldsymbol{\theta}$ .......... 13
Figure 3: Hazard functions associated with Log-logistic models. The shape of the Log-logistic hazard function is determined by its shape parameter, $\boldsymbol{k}$ .......... 14
Figure 4: More complex hazard function .......... 15
Figure 5: Illustrative example of a landmark survival model .......... 31
Figure 6: Illustrative example of a piecewise survival model .......... 35
Figure 7: Illustrative example of a cure model .......... 37
Figure 8: Illustrative example of isolated competing hazard functions over time .......... 52
Figure 9: Illustration of varying survival extrapolations from 10 simulated trial arms with the same underlying disease-specific hazard and using the same external data .......... 53
Figure 10: Disease specific survival and hazard functions, Scenarios 1-4 .......... 58
Figure 11A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 1 .......... 59
Figure 11B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 1 .......... 60
Figure 12A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 2 .......... 61
Figure 12B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 2 .......... 62

Figure 13A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 3 ........ 63
Figure 13B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 3 ........ 64
Figure 14A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 4 ........ 65
Figure 14B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 4 ........ 66
Figure 15: Convergence – percentage of simulations that failed to converge ........ 68
Figure 16: Bias in restricted mean survival time at 3-years, Scenarios 1 and 2 ........ 69
Figure 17: Bias in restricted mean survival time at 3-years, Scenarios 3 and 4 ........ 69
Figure 18: Bias in mean overall survival times (axis restricted to a maximum of +/- 5 years, Scenarios 1 and 2 ........ 70
Figure 19: Bias in mean overall survival times (axis restricted to a maximum of +/- 5 years, Scenarios 3 and 4 ........ 70
Figure 20: Plot of (a) survival and (b) hazard functions for various parametric and flexible parametric models up to the end of trial follow-up (3-years) for one realisation of Scenario 1 ........ 78
Figure 21: Plot of (a) survival and (b) hazard functions for various parametric and flexible parametric models extrapolated up to 40-years for one realisation of Scenario 1 ........ 78
Figure 22: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality up to the end of trial follow-up (3-years) for one realisation of Scenario 1 ........ 80
Figure 23: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality extrapolated up to 40 years for one realisation of Scenario 1 ........ 81
Figure 25: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality extrapolated up to 40 years for one realisation of Scenario 1. Background mortality is misspecified ........ 83
Figure 26: Plot of marginal survival for one realisation of Scenario 3, with varying follow-up lengths across the sub-panels. The grey shaded area shows the period where the survival functions is extrapolated (i.e. beyond the range of the trial follow-up). ........ 84
Figure 27: Plot of marginal hazard function for one realisation of Scenario 3, with varying follow-up lengths across the sub-panels. ........ 85
Figure 28: Plot of marginal hazard function for one realisation of Scenario 3, with varying assumed cure points ........ 86
Figure 29: Plot of (a) survival and (b) hazard functions for one realisation of Scenario 1. Piecewise models ........ 87

# 2. INTRODUCTION

## 2.1. BACKGROUND

### 2.1.1. *Survival analysis in economic evaluation*

Survival analysis is required to extrapolate from clinical trial data to estimate the lifetime costs and benefits of interventions and comparators in economic evaluations. To accurately model survival, models need to consider the hazards (rates of events) within the trial period and hazards beyond the trial.

Technical Support Document (TSD) 14 provides guidance on the process for selecting survival models, and explains the assumptions underpinning common, or “standard” parametric models[1, 2] (see list below for the models we include under this term). Each of the standard parametric models makes assumptions about the hazards over time:

- Exponential models assume that the hazard remains constant over time (Figure 1)
- Weibull models assume that hazards monotonically increase or decrease; that is there are no turning points and as time increases the hazard either consistently increases or decreases for larger values of time (Figure 1)
- Gompertz models assume that hazards monotonically increase or decrease, but the rate of change is assumed to be exponential (Figure 2)
- Log-logistic, log normal and Generalised Gamma models can represent hazards that monotonically decrease, or that initially increase and then decrease (one turning point) (illustrated for the log-logistic model in Figure 3). The Generalised Gamma can also represent hazards that initially decrease and then increase.

Therefore, the standard parametric models are limited with respect to the types of hazard that they can represent. This means that none of the standard parametric models can accurately model survival where there are two or more turning points, or where there are multiple important changes in the slope of the hazard function. More flexible models are required where hazard functions are observed, or expected - in the longer-term, to have complex shapes. TSD14 acknowledges this point, but provides relatively little detail on the more complex survival models that could be used.

**Figure 1: Hazard functions associated with Weibull and exponential models. The shape of the Weibull hazard function is determined by its shape parameter, $\gamma$.**

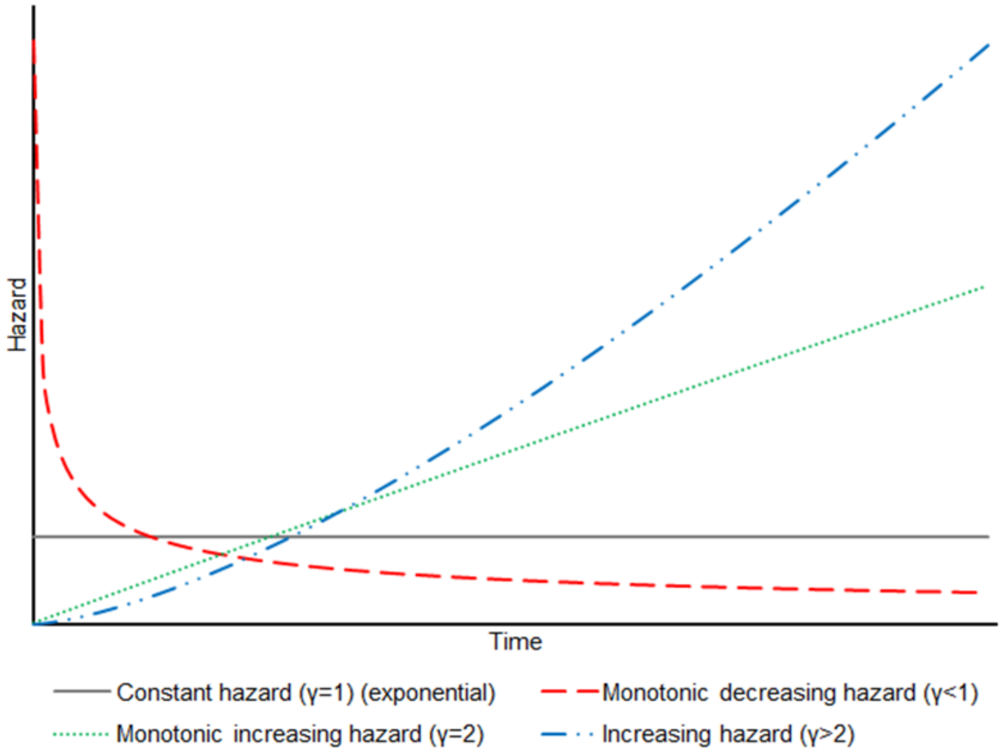


**Figure 2: Hazard functions associated with Gompertz models. The shape of the Gompertz hazard function is determined by its shape parameter, $\theta$**

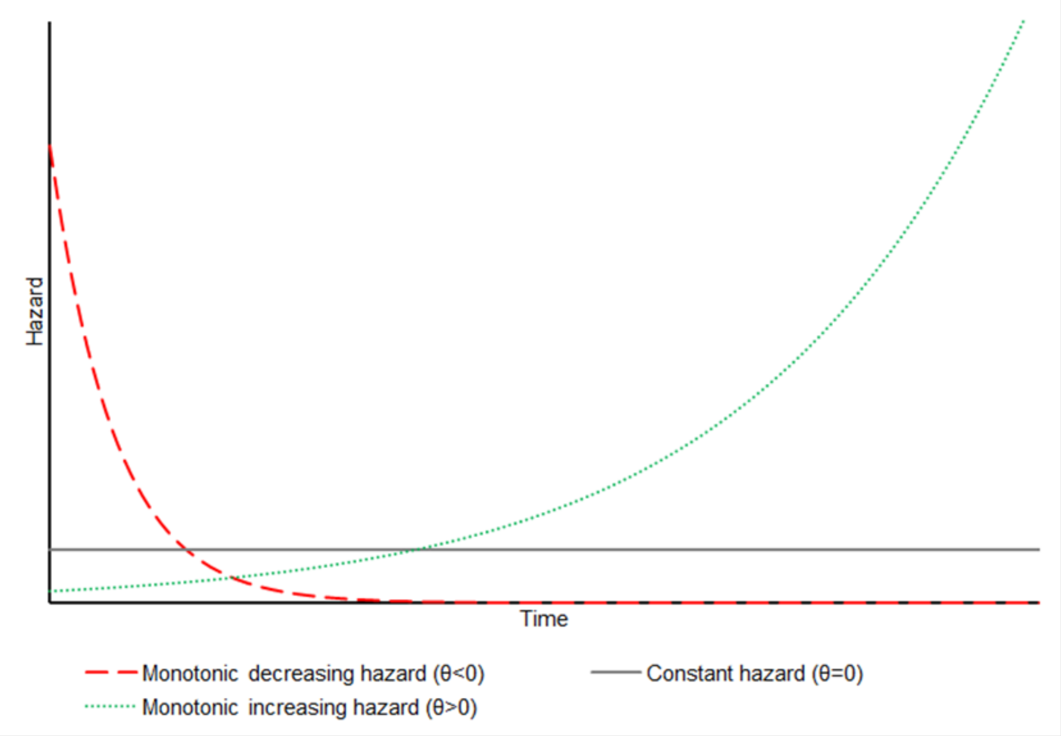
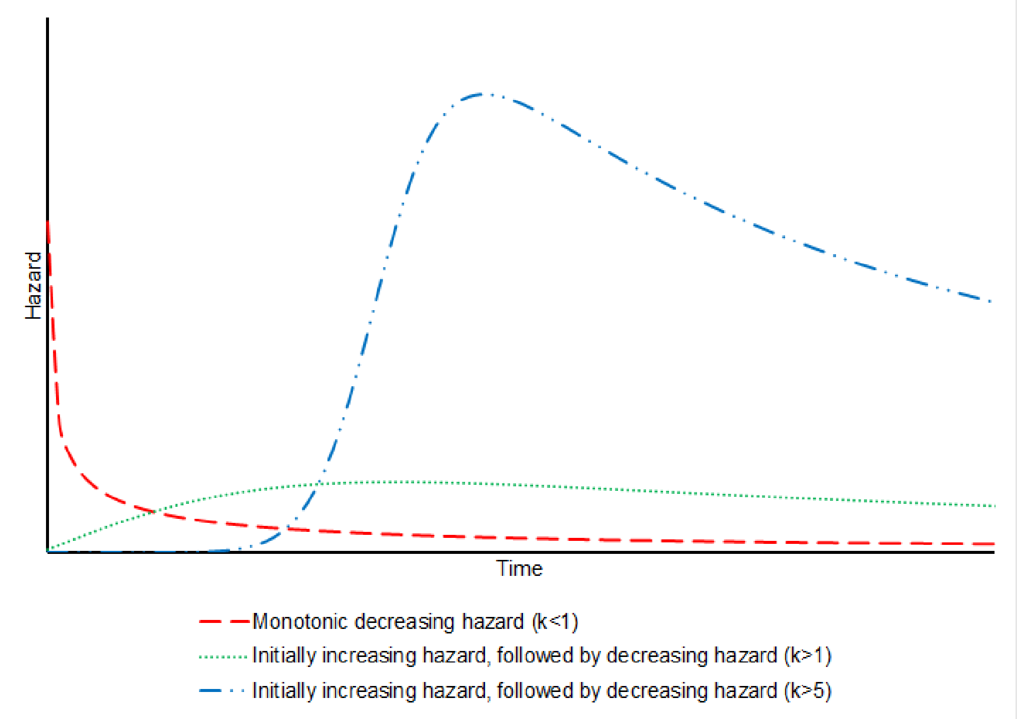


**Figure 3: Hazard functions associated with Log-logistic models. The shape of the Log-logistic hazard function is determined by its shape parameter, $k$**

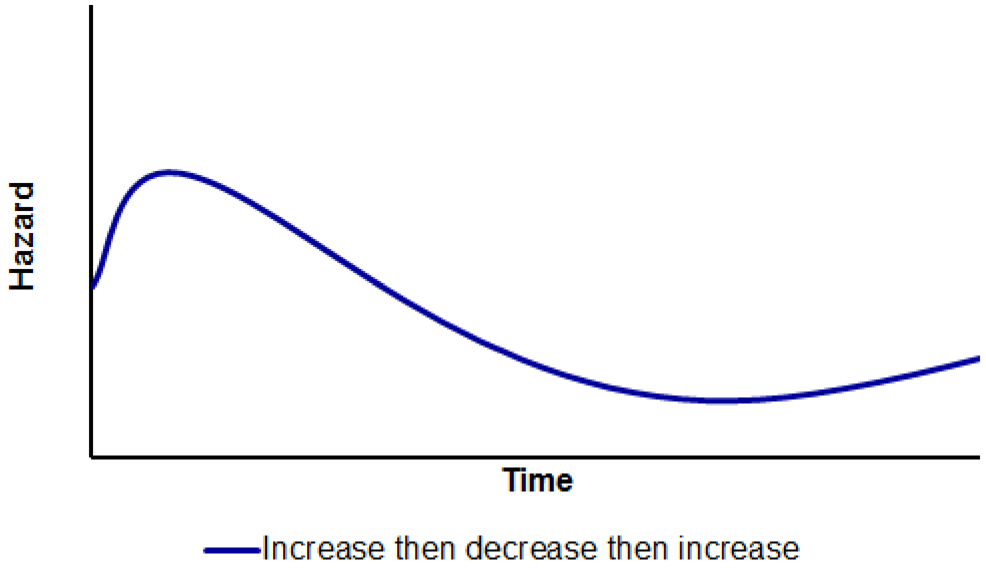


### 2.1.2. *Hazard functions*

Economic evaluations seek to model the cost-effectiveness of treatment for diseased populations in practice rather than only clinical trial populations, but clinical trials are usually considered to provide suitable data for informing survival models, especially with respect to relative treatment effects. Hence, it is usually desirable to use survival models that fit well to clinical trial data but which also extrapolate beyond the trial in a realistic way. Therefore, models that allow hazard functions with complex shapes may be advantageous.

The advent of immuno-therapy treatments for oncology has resulted in an increase in the use of complex survival models, because delayed responses to treatment and the existence of long-term survivors have been hypothesised to result in complex hazard functions[3-5]. However, complex hazard functions are not only conceivable in immuno-oncology. For instance, in most cancer trials the mortality (hazard) rate upon entry to the trial may be relatively low, due to trial eligibility criteria meaning that recruited patients must be fit enough to receive treatment with a potentially toxic (new) therapy. However, due to the nature of the disease, the mortality rate is likely to rise in the short-term. Then, over time, as the case-mix of the cohort changes because the sicker patients die, healthier patients and treatment responders survive and so the mortality rate decreases. In the longer term the effectiveness of the treatment might wane, or disease progression might occur, resulting in an increase in the hazard. Even if the treatment represented a cure for a small proportion of patients, in the very long-term hazards would be expected to rise, reflecting age-related mortality (Figure 4).

**Figure 4: More complex hazard function**



Longer-term changes in the hazard may not be observed within the trial period, but - given a realistic expectation that they will be observed beyond the trial period - these are relevant for inclusion in a model used for economic evaluation, where a lifetime time horizon is typically used. None of the standard parametric models could adequately represent the hazard function illustrated in Figure 4. It may be useful to consider survival models that can capture such hazard functions. Hazard functions are not routinely presented in NICE Technology Appraisals, but their inclusion may add to an understanding of the longer-term assumptions that are being made.

To determine whether survival models adequately represent the hazards observed in a trial, the model can be compared to the observed trial data. This is straightforward, and whilst important, often several models may be shown to provide a close fit to the trial data. The trial duration is usually very short in relation to the extrapolated period - that is, the period from the end of trial follow-up to the point at which all patients are expected to have died (if assuming a lifetime time horizon). It is the extrapolated period in which survival models typically diverge, often resulting in dramatically varying mean survival and cost-effectiveness estimates. The models which best fit the clinical trial data may not necessarily be the most appropriate for extrapolation. Determining which survival model provides the best representation of hazards beyond the trial period is challenging, as there are no trial data with which to compare. Consideration should therefore be given to whether the extrapolation is realistic, which may involve using external data sources, clinical expert opinion, or arguments around biological plausibility (incorporating knowledge around expected shapes of survival functions in given populations, or knowledge of specific disease subgroup characteristics).

Comparing various extrapolations of survival to external data sources represents one option for evaluating plausibility. An alternative approach would be to start with a pre-defined understanding of the various risks and mechanisms (such as competing risks, ageing effects, changes in case mix, and the waning effect of treatment), and their likely impact when governing the marginal survival. The model choice could then be consistent with, and directly model, these mechanisms, or this information could be used to define prior probability distributions in a Bayesian statistical analysis. Directly modelling these mechanisms, for example in a competing risk or multistate disease progression model, may require synthesis of evidence from multiple external sources and a Bayesian multi-parameter evidence synthesis approach may be appropriate in this setting[6]. We cover aspects of Bayesian approaches further in Section 2.8, and in our recommendations for future research.

### 2.1.3. *Use of flexible models in NICE Technology Appraisals*

The model selection algorithm presented in TSD14 advises when models that are more flexible than the standard parametric models may be required, but does not provide specific guidance on specific flexible models, their assumptions and limitations, or when they should be used. Flexible survival models are being increasingly considered in NICE Technology Appraisals. For instance, in TA498 (Lenvatinib with everolimus for previously treated advanced renal cell carcinoma) the evidence review group (ERG) used a flexible parametric model incorporating splines to model treatment duration[7]. In TA517 (Avelumab for treating metastatic Merkel cell carcinoma) the company used flexible parametric models with splines to model progression-free survival (PFS) and overall survival (OS)[8]. In TA478 (Brentuximab vedotin for treating relapsed or refractory systemic anaplastic large cell lymphoma) the company used mixture cure models for PFS and OS[9], and in TA463 (Cabozantinib for previously treated advanced renal cell carcinoma) the company used fractional polynomial models for PFS and OS[10].

These more complex modelling methods have not been used or interpreted consistently across appraisals. In TA483 the company, ERG and appraisal committee all interpreted a flexible parametric survival model with 2-knots as representing a model that implied that there were 3 heterogeneous subgroups within the patient population[11]. In fact, such a model simply represents a way of modelling a complex hazard function – it makes no assumptions about the number of heterogeneous subgroups directly. Such assumptions could be more justifiably associated with mixture models, which are different to spline-based flexible parametric models – however, even mixture models are usually interpreted simply as an alternative way for modelling complex functions. Inconsistent use and interpretation of complex survival models demonstrates the need for guidance explaining which flexible methods exist, how they work and what assumptions underpin them, and when they may (and may not) be appropriate.

### *This document*

Our aim is to describe different flexible survival models, and to use simulation studies and illustrative examples to demonstrate when and how these models can be used. Importantly, we also demonstrate the limitations associated with flexible (and standard) survival models, particularly with respect to extrapolation. Our intention is that companies preparing submissions and ERGs reviewing submissions can use this document to aid understanding of these models, which should improve the consistency with which they are used. This document can also be used by NICE technical staff and committee members to understand analyses performed. As mentioned, careful thought should be given to the biological and clinical justification to any statistical approach selected; the approaches detailed herein should not be considered as an extended list of survival methods to “try out” on data. Instead, care should be taken to think through the underlying mechanisms likely to be dictating short and long-term hazard/survival functions.

Survival analysis is particularly relevant in oncology where overall survival and progression-free survival are usually key components of the economic analysis, and where clinical trials are usually subject to a large amount of administrative censoring, with a short duration of follow-up in trials relative to the lifespan of patients. In this document our focus is on oncology, but the principles apply to other therapeutic areas.

We concentrate here on modelling and extrapolating the survival experiences of patients in a single arm of a trial. We do so because whether the effect of treatment is modelled as a relative effect (and applied to a baseline) or the two arms of a trial are modelled separately, the fundamental issues surrounding extrapolation over a longer time horizon remain the same.  However, it is still critical to both plot and justify the implied treatment effect when taking this approach. Whether modelling and extrapolating both arms separately, or  modelling the relative treatment effect directly, the long-term treatment effect should be considered, plotted for transparency, and justified.

In taking a single trial-based approach, we are discussing the potential to estimate the mean survival differences in the trial population. Should this differ from the target population of interest, then this must be factored into the estimation approaches for estimating the absolute measures of mean survival (see TSD18)[12, 13].

Research into methods for incorporating external information (such as from registry data, previous trials with longer follow-up, or elicited prior beliefs) to inform extrapolations is ongoing, and so it is not possible to make strict recommendations on this within this document. However, further information on incorporating background mortality and registry data into the survival modelling process will be discussed in Section 3. This is important for all models, particularly with respect to extrapolation since a review of NICE appraisals from 2011 to 2017 found only a minority undertook this[14].

In Section 2 we describe a variety of survival modelling approaches that can be used when hazard functions are complex, with respect to fitting survival models to observed data. In Section 3 we discuss extrapolation of survival functions beyond the trial period. Each modelling approach makes specific assumptions which are important to consider when using them to extrapolate survival functions - we provide detail on these and also discuss methods for incorporating external information within the survival modelling process. In Section 4 we present a simulation study to show the sensitivities of the different survival modelling approaches and to highlight the impact that they can have on survival estimates and population means.  In Section 5 we provide illustrative examples of some of the issues highlighted in the simulation study and discuss implications for practice. In Section 6 we provide discussion, recommendations, and suggestions for further research.

# 3. METHODS: FITTING TO OBSERVED DATA

## 3.1. Overview

In this section we describe a variety of survival modelling approaches that can be used when hazard functions are complex. We present the motivation behind each approach, their details with respect to formulae and assumptions, and their limitations. We do this with respect to fitting survival models to observed trial data. Given the typical sample sizes of RCTs and the corresponding number of events, consideration should be given to the number of parameters that can be reliably estimated from the available data for each of the described approaches. We compare the approaches in both a small and larger sample size in the simulation study conducted in Section 4.

## 3.2. Flexible parametric survival models

### 3.2.1. *Motivation*

Flexible parametric models (FPM) were developed because it was recognised that standard parametric models were often unable to capture adequately the underlying shape of hazard functions seen in applied studies[15]. Flexible parametric models use restricted cubic splines to enable hazard and survival functions with complex shapes to be accurately modelled. Restricted cubic splines are mathematical functions that can capture many complex shapes and thus enable more realistic hazard and survival functions to be estimated. The complexity of the function depends on the number and location of joining points of the function, with these joining points known as “knots”. The function is forced to be smooth by imposing constraints such that the function has continuous 1st and 2nd derivatives at the knots, i.e. the gradient and the rate of change of the gradient of the function also agree at the knots. Restricted cubic splines can be incorporated into any statistical model within a linear predictor by calculating derived variables known as *basis functions*. Increasing the number of derived variables (i.e. increasing the number of knots) results in a model that can take increasingly complex shapes.

### 3.2.2. *Details*

For some standard parametric distributions (Weibull, log normal, log-logistic) it is possible to transform the survival function to a scale which is a linear function of log time. For example, transforming the Weibull survival function, $S(t) = \exp(-\lambda t^{\gamma})$, to the log cumulative hazard scale gives,

*Equation 1*

$$\log[H(t)] = \log[-\log[S(t)]] = \log(\lambda) + \gamma\log(t)$$

This is a linear function of log time with intercept, $\log(\lambda)$, and gradient, $\gamma$ where $\lambda$ is the scale parameter and $\gamma$ the shape parameter of the Weibull distribution. Assuming a Weibull distribution can be too restrictive when, for example, there is a turning point in the hazard function. The assumption of linearity can be relaxed by using a more complex function of time. Flexible parametric survival models replace the linear function, $\log(\lambda) + \gamma\log(t)$ with a restricted cubic spline function of log time, $s(\log{(t)}|$ **γ**,$k_0$), where $k_0$ is a vector of knots and **γ** the associated parameters**.** Thus, the model becomes,

*Equation 2*

$$log[H(t)] = log[-log[S(t)]] = s(log(t)|\gamma, k_0)$$

This allows a much more flexible hazard function, which is able to capture a wide range of shapes. Covariates, $x$, with associated parameters, $\beta$, can be added to the linear predictor in Equation 2. For instance, a covariate for treatment group could be added.

*Equation 3*

$$\log[H(t)] = \log[-\log[S(t)]] = s(\log(t)|\gamma, k_0) + x\beta$$

The model represented by Equation 3 is a proportional hazards model and covariate effects can be interpreted as log hazard ratios – exponentiating the coefficients yields hazard ratios.

Alternative scales

When using FPMs the most common transformation of the survival function is the log(-log) transformation as described above. However, different transformations of the survival function can be used. For example, a logistic transformation of the survival function gives a linear function of log time for the log-logistic distribution and a probit transformation of the survival function gives a linear function of log time for the log-normal distribution. These extensions to relax the assumption of linearity are implemented in the same way as those described above.

The number and location of the knots.

A key issue when using FPMs is the number and location of the knots for the restricted cubic splines. The typical location in current software implementations is to place knots uniformly along the distribution of uncensored log event times with boundary knots placed at the minimum and maximum uncensored log event times. For example, with 5 knots (2 boundary and 3 internal), knots are placed at the 0th, 25th, 50th, 75th and 100th centiles of the uncensored log event times. Predicted survival functions within the range of the follow-up have been shown to be very insensitive to the number and location of the knots, provided that there are a sufficient number to capture the underlying shape[16-18]. However, consideration should be given to the typical sample size in RCT data, and, in particular, the number of events when trying to model complex hazard functions. Findings for large scale registry data have noted that having one or two more knots than necessary will have very little impact in terms of the predicted survival function within the range of follow-up, but it may be impractical to have many parameters for the spline function if there are few overall events. It should also be noted that when the aim is to extrapolate the survival function, then altering the number of knots may have an important effect on how the survival function is extrapolated beyond the data. Projection beyond the data is dictated by the hazard or survivor function estimated by the linear term (on a transformed scale) beyond the final knot, which may or may not be sensible when extrapolating (see Section 3).

The AIC (Akaike information criterion) and BIC (Bayesian information criterion) are often used as an informal guide for selecting the number of knots, with the minimum value of the AIC or the minimum value of the BIC giving the “best” fitting model. However, this is only a guide and if one wants to understand the (lack of) impact of choosing a different number of knots, then sensitivity analysis is recommended. Similarly, it is worthwhile emphasizing that the AIC/BIC may help in selecting a model that fits the data within the length of follow-up, but they provide little information about how well the model extrapolates to longer time points. A “good fit” within the range of the data may nevertheless lead to implausible extrapolations (see Section 5.2); for instance, a survival function that reaches a plateau that lasts for many years, resulting in a hazard function that would be below that of the general population.

Time-dependent effects

As stated in TSD14, parametric models can be fitted with the treatment indicator as a covariate (thereby assuming a proportional treatment effect), or can be fitted independently to each treatment arm (not assuming proportional treatment effects). Assuming proportional treatment effects is restrictive and may result in poorly fitting (and implausible) survival models and extrapolations. One of the advantages of FPMs is the ease with which the proportional hazards assumption can be relaxed. In general time-dependent effects are implemented by creating additional spline terms and creating interactions between these and the covariate of interest (such as the treatment indicator). For example, with just a single covariate *x*, a model relaxing proportional hazards is given in Equation 4,

*Equation 4*

$$\log[H(t)] = \log\big[-\log[S(t)]\big] = s(\log(t)|\gamma, k_0) + x\beta + x\, s(\log(t)|\delta, k_1)$$

In Equation 4 the first spline function is the baseline hazard function, but now the treatment effect is also a function of time given the second spline function. Note that there can be a different number of knots for the baseline and the time-dependent effect. This is often sensible as the shape of the underlying hazard function is usually more complex than the deviation between two hazard functions[19]. Alternatively, separate models could be fitted for each treatment group. This is effectively the same as fitting a complex interaction between the treatment covariate and follow-up time. When separate models are fitted for each treatment group, implausible effects of the treatment effect (hazard ratio) beyond the range of follow-up could be projected. Plotting the extrapolated hazard functions for both treatment groups together with the implied hazard ratio helps demonstrate what is inherently being assumed about the relative treatment effect.

In summary, FPMs assume that the spline function is adequate to model the hazard function within the range of follow-up. This depends upon having sufficient knots. It is often assumed that the effect of any covariates act proportionally on the baseline hazard rate, but this can be relaxed (allowing non-proportional hazards) through interaction effects with a function of time (or through fitting separate models).

### 3.2.3. *Limitations*

The use of FPMs is associated with a number of limitations:

- The number of parameters used to model the hazard function is decided by the user. The AIC and BIC can be used as a guide, but on a single dataset it's unclear whether the most sensible model has been chosen. The AIC/BIC should not be used as the sole basis for model selection when the aim is to extrapolate (this can be said for any parametric model).
- FPMs will generally provide extremely good fits within the range of the observed data, given a sufficient number of knots have been used, but this does not mean that their extrapolations will be reliable.
- If external data are not incorporated then extrapolation associated with an FPM is based completely on the linearity assumption (on a transformed scale of the survival function), which may result in implausible projections. With large numbers of knots the extrapolation may be based upon the trend towards the end of follow-up, which may be based on a limited number of events.
- With multiple time-dependent effects the hazard ratios for one covariate is dependent on the value of another time-dependent covariate. This is a problem when trying to summarise many covariates with time-dependent effects using hazard ratios. This is because the hazard ratio for one time-dependent effect is dependent on the values of a second time-dependent effect, even when there is not an interaction between the covariates. However, when one is interested in prediction of survival functions (and estimating the mean survival or restricted mean survival time [RMST] for use in an economic decision model) then this is not an important limitation.

### 3.2.4. *Further Issues*

Flexible parametric models are usually fitted on the log cumulative hazard scale. One reason for this is that the hazard and survival functions are analytically tractable which enables very quick estimation. However, it is also possible to fit models on the log hazard scale and again use restricted cubic splines to model the effect of time. These models require numerical integration to obtain the cumulative hazard during estimation of the model parameters.  If making assumptions about the effect of a covariate (such as treatment) after the end of follow-up, then it may be easier to think about the plausibility of assumptions on the log hazard scale rather than the log cumulative hazard scale. For example, if we want to assume that the hazard rates associated with two treatment groups are the same after a certain point in time, this is straightforward on the log-hazard scale, but would mean imposing that the gradient of the treatment groups was the same on the log cumulative hazard scale[20].

There has been some work using penalized spline functions[21]. In these models a larger number of knots are chosen and then a penalty function incorporated into the likelihood to force the function to be smooth. However, the user still needs to define the number of initial knots and the type of penalty function to incorporate. It is unclear what the impact on extrapolation would be using such an approach.

### 3.2.5. *Other flexible parametric models*

A poly-hazard model[22] represents another type of flexible parametric model. These are distinct from mixture models, which are described in section 2.3. A poly-hazard model assumes an overall hazard function which is the sum of K hazards. This can be thought of as a cause-specific competing risks model (that is, each component can be thought of as contributing additively to the overall hazard). Specifying a sum of hazard functions allows the overall hazard to provide a far more flexible function compared to the standard parametric models often used in practice. Each component can have a different functional form, for example if we have three components, then we may assume one exponential, one Weibull and one Gompertz, and hence covariates can also be modelled differently for each hazard. As an example, Demiris et al[23], proposed a poly-hazard model defined as follows,

*Equation 5*

$$h(t) = \sum_{k=1}^{K} h_k(t),$$

where $h_k(t)$ is the hazard function for the k$^{th}$ mixture. Demiris proposed the use of Weibull models for all hazard components, motivated by wanting to capture a bathtub hazard function, where,

*Equation 6*

$$h_k(t) = \lambda_k \gamma_k t^{\gamma_k - 1}$$

Each of the shape and scale parameters can have their own linear predictor (on the log scale in this case) which can include covariate effects. There is no restriction for each hazard component to be specified as a Weibull, indeed any of the standard parametric distributions could be used, including different distributional choices for different hazard components.

A poly-hazard model might be appropriate if it is believed that there are multiple competing causes of death with different hazard trajectories, and if some information about these trajectories is available. This is similar in principle to the setting described in section 2.7, where known population mortality rates are used as a fixed hazard function, and a separate additive hazard function is modelled in excess of the background mortality.

## 3.3. MIXTURE MODELS

### 3.3.1. *Motivation*

Mixture models may be appealing if it is believed that different clusters or sub-populations of patients have different hazard and survival profiles. For example, people who achieve a response to treatment may have a different survival profile to those who do not respond. In this sense, a mixture model is sometimes interpreted as accounting for the fact that different sub-populations within a trial have different survival profiles, which are represented by the different survival distributions included in the mixture model. However, mixture models assign a probability to each patient of being in each distribution included in the mixture – patients are not definitively segregated into separate groups. In parallel to this, a mixture model may also be thought of as a way of specifying a flexible hazard trajectory (similar to the motivation behind flexible parametric models above), in other words it simply represents an approach for modelling complex hazard functions, which involves using a mixture of parametric distributions [24].

### 3.3.2. *Details*

A mixture model can be defined as follows,

*Equation 7*

$$f(t) = \sum_{k=1}^{K} p_k f_k(t), \quad \text{where } \sum_{k=1}^{K} p_k = 1,$$

where $f(t)$ is the overall distribution function, made up of additive component distribution functions, $f_k(t)$, for the k<sup>th</sup> mixture, and $p_k$is the proportion that the k<sup>th</sup> mixture contributes to the overall distribution function. There is no restriction for all components to have the same distribution, indeed any of the standard parametric distributions could be used, including different distributional choices for different mixture components. Since each mixture component can be adjusted for covariate effects, in all parameters, including the mixture probabilities, they can be used to identify subpopulations which may respond differently to treatment. To ensure the constraint that $\sum_{k=1}^{K} p_k = 1$, we may model the mixture probabilities with a multinomial distribution[25].

Standard model selection criteria such as AIC and BIC can be used as a guide to select the number of mixtures, and the distributional form for each contributing distribution function. However, when extrapolating beyond the trial period it is important to note that these extrapolations will be driven by the combination of both the weights received by each mixture component and the relative magnitude of their respective hazard rates.

Mixture models make the following assumptions:

- The observed (and expected) hazard function can be appropriately modelled using a mixture of standard parametric models.
- The data within each mixture component is sufficient for robust survival modelling, and the distributional choice for each mixture component is valid.

- If a mixture model is used to represent latent subgroups, we cannot assign patients to subgroups with certainty, we can only estimate the probability that each individual is in each group.
- An extrapolation from a mixture model will be driven by the combination of both the weights received by each mixture component and the relative magnitude of their respective hazard rates, and as such, these should be assessed for their plausibility for long term extrapolations, using either prior knowledge or longer-term survival information from other sources.

### 3.3.3. *Limitations*

Mixture models are subject to the following limitations:

- Choosing the number of mixture components.
- Choice of each mixture distribution.
- If the goal is to capture a complex hazard function, then arguably this is much easier to do with a flexible parametric modeling framework using splines (as described and discussed in Section 2.2).
- They are at risk of over-interpretation, such as, concluding that the sample of patients is made up from *n*-mixtures with different risk profiles. Despite the mixture components, we obtain an overall (potentially complex) hazard function which may be equally as well estimated with an alternative, more easily implemented and parameterised, approach - therefore mixture models may give misleading conclusions.
- Challenge of model convergence (see Section 4). Mixture models are notoriously difficult to estimate, given the inherent problem of multiple maxima[24]. For example, in a two-component model where $p_1 + p_2 = 1$, there are two combinations of the same parameters which would give an identical function. In other words, if $p_1 = 0.2$ and $p_2 = 0.8$, given the estimated component distribution parameters, we would get the same model when $p_1 = 0.8$ and $p_2 = 0.2$, with the component distribution parameters swapped. In order to estimate each mixture component and hence an overall model, there must be sufficient number of events.

### 3.3.4. *Further Issues*

Each parameter in a mixture model could be adjusted for treatment, including each component's shape and scale, for example, and indeed the mixture parameters themselves. Alternatively, completely different mixture models could be fitted for each treatment group.  Alternatively, an overall shared treatment effect may be estimated, in which case the mixture components are used simply to model a more flexible baseline function. In any case, mixture models must be interpreted with care – any reference to the nature and presence of sub-populations should be considered carefully, with reference to the mixture probabilities assigned to each patient and covariate effects if they are included. In addition, the existence of sub-populations with different survival profiles should be based on biological plausibility.

## 3.4. LANDMARK MODELS

### 3.4.1. *Motivation*

The motivation behind landmark models is to use a modelling approach that acknowledges that the survival experience of a patient might be substantially different depending upon whether or not an individual responds well to treatment. It is therefore assumed that response represents an important surrogate for survival. Landmark models use a defined "landmark" time point, at which point patients are split into groups according to their response category. Typically RECIST criteria are used, categorising individuals into "complete response", "partial response", "stable disease" and "progressive disease" groups, or these categories are merged into "response" and "non-response"[26]. Separate survival models are fitted to each response group, from the landmark time-point. Survival for the whole population beyond the landmark time-point is then estimated by weighting the survival function for each response group by the proportion of patients within that group. Prior to the landmark time-point the Kaplan Meier survival function can be used to estimate survival, or a parametric model could be used. Theoretically, the survival models for each group can take any form, but standard parametric models tend to be used[27, 28].

Because landmark models allow different survival models to be fitted to each response category, they are able to represent complex hazard functions with turning points. For instance, the model used for the non-response group may have a high or increasing hazard and the model for the response group may have a low and decreasing hazard. In this case, over time, as non-responders die, the hazard for the remaining population will decrease. With the right combination of models, turning points (possibly multiple turning points) could be represented.

Landmark models have been used in NICE TAs. In TA421 (Everolimus with exemestane for treating advanced breast cancer after endocrine therapy)[29] the ERG used a landmark model, applying exponential models to each response group.

### 3.4.2. *Details*

The survivor function beyond the landmark time using the landmark approach can be described using

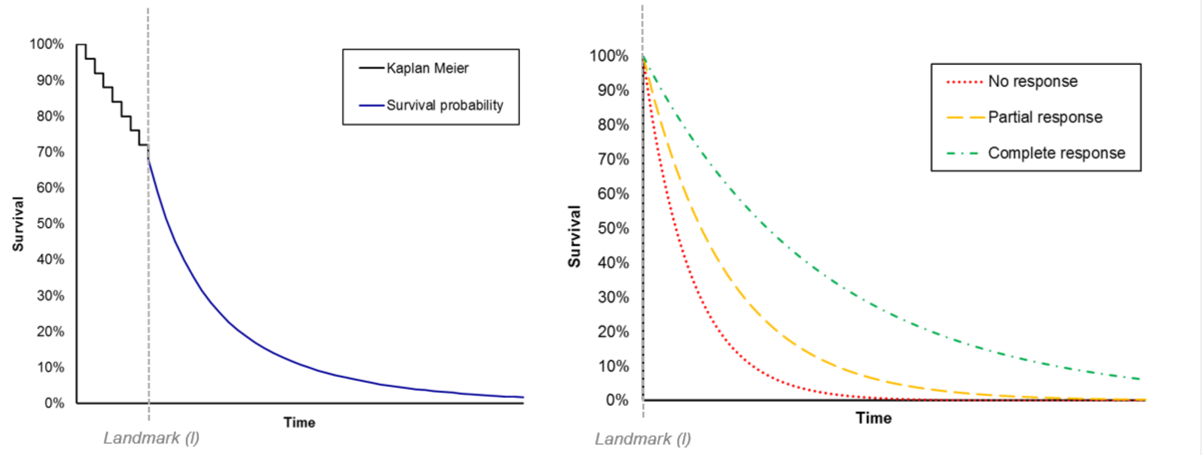
*Equation 8*

$$\text{Survivor function: } S(t) = S(l) \times (\sum_{i=1}^{k} S_i(t|T > l) \times \frac{n_i}{\sum_i^k n_i})$$

Where $S(t)$ is the survival at time $t$, $l$ is the landmark time point, $S_i(t|T > l)$ is the survival at time $t$ given survival to time $l$ for patients in the $i$th response category, $k$ is the number of response categories, and $n_i$ is the number of patients in response category $i$. The $S(t|T > l)$ will be dictated by the survival model used for each response category, and are conditional on the landmark time point, $l$.

A landmark model is illustrated in Figure 5. In the illustration, survival up until the landmark time point is based upon the Kaplan-Meier curve. After that it is estimated using a combination of models fitted to three response groups, which are combined based upon the proportion of patients in each group.

**Figure 5: Illustrative example of a landmark survival model**



Survival probability at time (t) =

Survival probability at landmark × ( % patients with no response × survival probability at time (t) for no response + % patients with partial response × survival probability at time (t) for partial response + % patients with complete response × survival probability at time (t) for complete response )

Landmark models make the following assumptions:

- The "landmark" time point for assessing response is appropriate. It is not always straightforward to determine what the landmark time-point should be because response may be measured at multiple time-points during the study.
- In combination with the landmark time-point, the response categories are clinically meaningful and represent a surrogate for survival.
- There are a sufficient number of events in each group for robust survival modelling, and the survival model for each group appropriately capture the long-term survival functions.

### 3.4.3. *Limitations*

Landmark models are subject to the following limitations:

- Landmark time-points may be arbitrary, and may importantly influence the results of the analysis. An early landmark time-point may miss delayed responses, whereas a late landmark time-point may result in less meaningful categorisation as a proportion of patients (likely to be non-responders) may die before the landmark point is reached. This will have important consequences in terms of the estimation of uncertainty around the estimates also.

- If, in combination with the landmark time-point, response categories are not considered to represent good surrogates for survival, it may not be justifiable to estimate survival separately for the categorised groups.
- Splitting the sample into response categories means that the size of each group can be small (with a consequent small number of events for each category), leading to large standard errors and uncertainty when fitting survival models. If the landmark time-point is relatively late, this may be a particular problem in the "no response" group, as many of these patients may have died before the landmark time. In addition, there may be very few deaths in "good response" groups, which may make fitting robust survival models problematic.
- Response may not be measured in all patients, or could be subject to error if it is measured.
- If the survival models used for the response categories are not appropriate to capture the shape of the specific hazard function, then the overall projection may be inappropriate also.

### 3.4.4. *Further issues*

Landmark models allow complex hazard functions for the trial population to be represented even if standard parametric models are used to model survival for each category. Typically, standard parametric models *are* used to model survival for each category within a landmark model. However, this assumes that standard parametric models can appropriately represent the hazard functions within each category, which may not be the case. This is not a limitation of the modelling approach *per se*, because more flexible survival models could be used within a landmark approach. However, it is a limitation of the way the landmark modelling approach is generally applied and is particularly relevant when considering the extrapolated portion of the survival function, where additional changes in the hazard might be expected (see Section 3).

## 3.5. PIECEWISE MODELS

### 3.5.1. *Motivation*

The motivation to use piecewise models may be under circumstances where standard parametric models have not appeared to provide a good fit to the data, or where multiple sources of data exist for different time periods. When piecewise models are fitted to just one dataset, the time-points for the cuts are often explicitly chosen by the user, although approaches to allow random changepoints have also been developed[30, 31].

One approach to piecewise modelling would be to examine the hazard function over time and make a judgement as to whether, and if so where, sufficient changes in the hazard occurred that require a new survival model to be fitted. This decision-making could also be made a-priori by considering the biological mechanisms governing the complexity of the hazard function over time. For instance, in a given context there may be an important change in the hazard at 3 months and at 6 months, and so separate survival models are fitted to the 0-3 months, 3-6 months, and 6+ month time periods. Theoretically, the survival models for each time period can take any form, but standard parametric models tend to be used. By using different survival models for each time period, flexible hazard functions over time can be represented, even if standard parametric models are used for each segment of the survival function[32, 33]. For instance, the initial survival model may have an increasing hazard, the second survival model may have a decreasing hazard, and the third may allow hazards to increase again.

Piecewise models have been used in several NICE TAs. For instance, in TA490 (Nivolumab for treating squamous cell carcinoma of the head and neck after platinum-based chemotherapy)[34] piecewise models were fitted to segmented overall survival, progression-free survival and time-to-treatment discontinuation data. Piecewise log normal models were fitted to overall survival data and piecewise generalised gamma models were fitted to segmented progression-free survival and time-to-treatment discontinuation data. Piecewise models have also been used in TA387 (Abiraterone for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated)[35] and TA421 (Everolimus with exemestane for treating advanced breast cancer after endocrine therapy)[29].

In some piecewise models, the Kaplan Meier survival function is used to represent the initial section of the survival function and an exponential function is adjoined to a pre-determined point of the Kaplan Meier. This approach has been popularised by the

Liverpool Reviews and Implementation Group (LRiG), and as such has become known as the “Liverpool approach”[36, 37]. Due to the limitations associated with extrapolating using an exponential model (i.e. with a constant hazard rate), other researchers have adjoined different survival distributions to pre-determined points of the Kaplan Meier survival function. For instance, in TA391 (Cabazitaxel for hormone-relapsed metastatic prostate cancer treated with docetaxel)[38] a Weibull model was adjoined to a pre-determined point of the Kaplan Meier survival function, and in TA519 (Pembrolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum-containing chemotherapy)[39] a log normal model was attached to a pre-determined point of the Kaplan Meier survival function.

### 3.5.2. *Details*

The overall survivor function of the piecewise modelling approach can be described as follows. Consider partitioning time into $J$ intervals, with cutpoints $0 = t_0 < \mathrm{t}_1 < \cdots < \mathrm{t}_\mathrm{J} = \max(\mathrm{t})$. The hazard component for the $j^{th}$ interval: $\lambda_j(t)$ for $t$ in $[t_{j-1}, t_j)$, can be defined by an exponential (constant), or other parametric form. The overall baseline hazard function, $\lambda_0(t)$, then is defined separately depending on the intervals, taking the value of $\lambda_j(t)$ for times in the $j^{th}$ interval. The overall survival function then relates to the summed cumulative hazard function over time, $\Lambda_0(t)$:

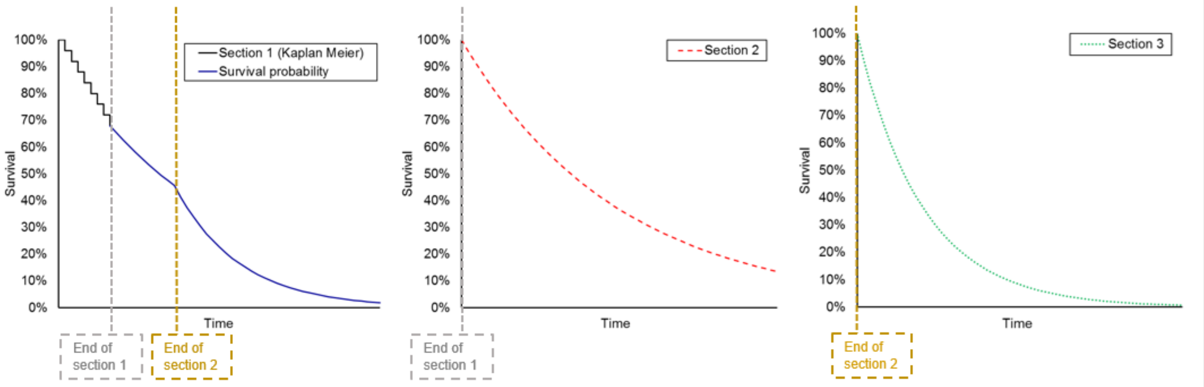
*Equation 9*

$$\text{Survivor function: } S(t) = \exp\left(-\Lambda_0(t)\right)$$

This could also be equivalently described in terms of the multiplication of conditional survival functions defined by the various parametric forms for each of the $J$ intervals, as in the illustration below.

A piecewise survival model is illustrated in Figure 6. In this illustration the initial part of the survival function is based on the Kaplan-Meier survival function. Different survival models are then fitted to sections 2 and 3 of the curve and are then adjoined to create a complete survival function.

**Figure 6: Illustrative example of a piecewise survival model**



Survival probability at time (t) =

Survival probability at time (t) in Section 3 × (Survival at End of section 1 × Survival at End of section 2)

Piecewise models make the following assumptions:

- The point(s) at which new models are fitted is/are appropriate
- The survival model used for each section is appropriate
- The data within each section are sufficient for robust survival modelling

### 3.5.3. *Limitations*

Piecewise models are subject to the following limitations:

- If the cutpoints are selected through visual inspection of the hazard or survival function then assessing a sufficient change for a new cutpoint may be difficult. Therefore, the cut-points for the various intervals may be arbitrary and may importantly influence the results of an analysis[40].
- Piecewise models may appear clinically unjustifiable and implausible, if sudden changes in hazards are modelled (i.e. with a discontinuity in the hazard function). For instance, if the hazard function for the survival model illustrated in Figure 6 was plotted it would show substantial “jumps” in the hazard at the join points for the different survival models. Methods could also be applied to allow continuity of the piecewise hazard function at the changepoints; these are similar in principle to otherwise methods for flexible smoothing the hazard function such as the use of splines (see Section 2.2).
- Where a piecewise model is fitted to a single dataset, splitting the data into sections according to time means that sample sizes are reduced in later

segments of the curve. This is a particular issue in later sections of the curve, where patient numbers at risk may be very small and the number of observed events may be low, leading to large standard errors and uncertainty when fitting survival models. A key point is that it is the model fitted to the latest section of the curve that is used for extrapolation.

- If the survival models used for each section are not appropriate, the overall projection will be adversely impacted. This is particularly important for the final section.

### 3.5.4. *Further issues*

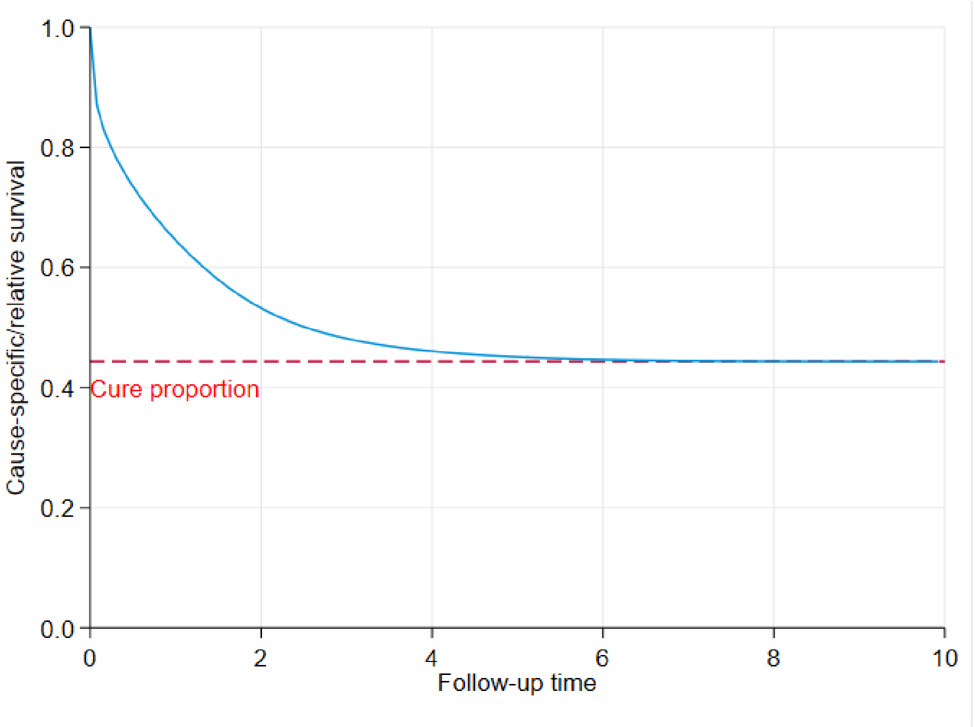
Piecewise models are usually implemented using standard parametric models applied to segments of the survival data. Whilst this allows complex observed hazard functions to be accurately represented it is not necessarily sufficient if additional changes to the hazard are expected beyond the period for which data are available (see Section 3). This is not a limitation of the piecewise modelling approach because a more flexible model could be used within a piecewise approach, or external data could be used within a piecewise model. In fact, this has been seen in NICE TAs - for example in TA268 (Ipilimumab for previously treated advanced (unresectable or metastatic) melanoma)[41] the company used a piecewise model for overall survival that adjoined a Gompertz model to the Kaplan-Meier curve at the 18 month time-point, and then used hazard rates from a registry study from year 6 onwards. However, thought must be given to the covariate profile of individuals still at risk at each timepoint when defining the appropriate hazard function from the external data, which we discuss further in Section 3.

## 3.6. CURE MODELS

### 3.6.1. *Motivation*

Traditionally, cure models have been used in situations where a proportion of individuals will never experience the event of interest (often used with disease-specific deaths as events). In such a situation, after a certain point in time, there will no longer be individuals having the specific events. This means that the hazard rate will be zero and the survival function will have a plateau at a non-zero value of the cause-specific survival function (see Figure 7).

**Figure 7: Illustrative example of a cure model**



Clearly, when using a cure model in the context of human survival, the event cannot be all-cause mortality as the survival function must reach zero. Thus, cure models are often fitted using cause-specific survival as the event of interest (with other cause mortality also modelled), or in an excess mortality/relative survival framework (see Section 2.7). Cure models may be attractive in the context of treatments for cancer, if it is believed that a proportion of patients will not die from their disease. In this scenario, a cure model may be used to estimate the cure fraction, and to estimate survival for uncured patients. It is important to note that cure is not defined at an individual patient level, but at a population level i.e. that the overall cause-specific hazard diminishes to zero). By combining the hazard function of the uncured fraction with the hazard function of the cured fraction cure models are able to estimate overall hazard functions that have a complex shape.

### 3.6.2. *Details*

<u>The mixture cure model</u>

The most common type of cure model is the mixture cure model[42, 43]. This model considers there to be two groups of individuals, those cured of their disease and those who are uncured.

*Equation 10*

$$S(t) = \pi + (1 - \pi)S_u(t)$$

where $\pi$ is the proportion of cured patients, $(1 - \pi)$ is the proportion of uncured patients and $S_u(t)$ is the survival function of the uncured patients. It is important to remember that if this is used in a cause-specific setting, neither $S(t)$ or $S_u(t)$ would give a real world probability and to obtain these other cause mortality would need to be taken into account. That is, careful consideration of the competing risks and the cause-specific hazard from causes due to the disease of interest and causes other than the disease of interest are required to obtain the marginal all-cause survival, with careful accounting of factors that influence both mortality rates.

An alternative to the cause-specific mortality setting is to incorporate expected survival directly into the model and thus fit a relative survival model. In order to do so, information on the general mortality rate (stratified by characteristics such as age and sex) is needed for the event time for each patient to offset the all-cause mortality model. The event now becomes death from any cause and our definition of cure changes to when the mortality rate amongst the diseased patients returns to the same level as that expected in the general population. We incorporate expected survival as follows:

*Equation 11*

$$S(t) = S^*(t)(\pi + (1 - \pi)S_u(t))$$

where $S^*(t)$ represents the expected survival function. These models have been applied in a population-based cancer setting and thus tend to be fitted to much larger datasets than generally seen in randomised clinical trials[44, 45]. In small datasets there may be issues around the practicality and plausibility of being able to reliably estimate the cure fraction.

The non-mixture cure model

An alternative type of cure model that uses an alternative mathematical function to define an asymptote for the survival function (a point that is approached as time tends to infinity – see example of the cure fraction in Figure 7) is the so called non-mixture model which is defined as:

*Equation 12*

$$S(t) = \pi^{F_z(t)}$$

where $F_z(t)$ is a cumulative distribution function. As with the mixture model, this usually takes a parametric form, usually using simple parametric distributions such as the Weibull or log-normal distribution. The non-mixture model can be extended to the relative survival setting in a similar way to the mixture model by incorporating expected survival information. In addition, the non-mixture model can be made even more flexible through defining the cumulative distribution function using restricted cubic splines, through extension of the flexible parametric survival models[44] described in Section 2.2. When using a flexible parametric cure modelling approach, the time at which cure is assumed (i.e. the point at which the disease-specific survival is assumed to reach zero) can be fixed by the user through the placement of the final boundary knot – in this setting the splines are calculated in reverse order and extra constraints are placed on the final parameters. This allows a flexible approach by which one can assume the effect on disease-specific mortality is effectively diminished to 0, and allow from that point onwards the all-cause hazard to be defined completely by other-cause mortality (potentially using an external data resource, such as registry data or general population mortality rates). We cover this concept with an example in Section 5.7.

Cure models make the following assumptions:

- Data available are sufficient to reliably estimate a cure fraction.
- A cure fraction exists, and cure is a reasonable assumption at a given time point (this depends on timescales, and to some extent the method taken). Remembering that this is a population cure level i.e when the disease-specific hazard diminishes to 0.
- The distribution chosen to model the non-cured fraction is appropriate.

### 3.6.3. *Limitations*

Cure models have the following limitations:

- In order to reliably estimate the cure fraction we need sufficient numbers at risk in the tail of the distribution. After a certain point in follow-up, we would expect there to be no further (cause-specific) events. If the number at risk towards the end of follow-up is low, which is common in randomised trials, then it is highly questionable as to whether it is sensible to impose such a strong assumption as cure. Assuming cure when it is not a realistic assumption could lead to very poor extrapolations.
- Although it is common to use measures such as the AIC and BIC to select the distribution to model for the uncured patients, one should be particularly cautious when fitting cure models. This is because when fitting cure models there is particular interest in the tail of the distribution where there are not many individuals at risk and not many events. The AIC and BIC will give more weight to how the model fits at the start of follow-up (as this is typically where the density of events will be) with very little weight towards the end. Due to their different shapes it is possible that different standard parametric models used to model the uncured fraction will result in very different cure fractions being estimated, therefore it is very important to attempt to select an appropriate model for the uncured fraction, based on biological plausibility.
- If the cause-specific survival function does reach close to a plateau then any reasonable approach to capture the shape of the hazard should estimate a cause-specific hazard function close to zero and thus estimated survival functions should be similar whether cure is assumed or not. However, cure models can be useful to fix the point at which disease-specific mortality based on the trial is assumed to no longer impact - although this time point must be clearly justified (see Section 5.7 for examples of this). Many standard cure models do not fix this timepoint, and the cure fraction is assumed to be reached at infinity.

### 3.6.4. *Further issues*

Cure models are a tool to force a plateau in a survival function. This plateau cannot be long-term for all-cause survival, so thought should be given to the outcome being modelled when considering cure. Within the range of follow-up, cure may appear reasonable because of limited sample sizes, and a low, but non-zero mortality rate - this should be considered particularly in the RCT setting. Cure models can seem attractive in some clinical settings - but care then needs to be given to incorporating background or other cause mortality when estimating longer-term survival.

## 3.7. EXCESS MORTALITY / RELATIVE SURVIVAL MODELS

### 3.7.1. *Motivation*

Excess mortality models are typically applied in population-based cancer registry data when information on cause of death is either missing or considered unreliable[46]. To our knowledge, this approach has not been used in health technology assessment. The concept behind these models is to isolate the cause-specific mortality by partitioning the all-cause mortality into that due to other causes and the excess mortality caused by the disease of interest. A similar approach could be taken in a clinical trial setting or else a cause-specific model could be fitted instead using cause of death information that may more reliably be recorded in a more controlled setting. In either case, a parametric model can then be applied to the isolated excess/cause-specific mortality. This approach is typically used to obtain covariate effects relating to the cause of interest, but can also be used to conduct a competing risks analysis. This may be particularly useful when making long-term extrapolations as the patterns of the disease-specific mortality and the other cause mortality are likely to be very different over time. Hence, modelling these separately and combining them to give estimates of long-term all-cause survival may be appealing. By separately modelling disease-specific mortality and with the potential to re-introduce estimates of long-term other-cause mortality, relative survival approaches are able to estimate overall hazard functions with complex shapes. This is synonymous to the poly-hazard setting introduced in Section 2.2.5, but here a fixed population mortality rate is used, with no uncertainty to reflect the mortality rate for other causes. The excess mortality approach was also introduced and motivated in the previous section (Section 2.6) when introducing the concept of population cure in Equation 10. In the cure setting, we place a constraint on the long-term relative survival function to reach a plateau.

### 3.7.2. *Details*

The all-cause mortality rate can be broken into two constituent parts:

*Equation 13*

$$h_i(t) = h_i^*(t) + \lambda_i(t)$$

where $h_i(t)$ is the all-cause mortality, $h_i^*(t)$ is the background mortality typically obtained from population mortality rates stratified by age, sex and calendar year (and other general determinants of population mortality rates) and $\lambda_i(t)$ is the excess mortality rate. Equation 13 can be transferred to the survival scale and rearranged to give the following relation:

*Equation 14*

$$R_i(t) = \frac{S_i(t)}{S_i^*(t)}$$

Meaning that the relative survival is the ratio of the all-cause survival and the expected survival in the background population.

It is possible to adapt parametric models to this setting and fit for instance a flexible parametric excess mortality model; or in the case of cause of death information any standard parametric model could be fitted to the cause-specific data, but this requires accurate cause of death information in order to partition the all-cause hazard function appropriately. The long-term extrapolations would then rely upon extrapolated excess mortality, and a re-introduction of the background mortality rates to capture other-cause mortality. In the short timeframe of a typical RCT, the excess mortality rate is often similar to the all-cause mortality rate, as many of the deaths in the short-term will be associated with the disease under study, and the mortality rate will be higher than in the general population. In isolating and modelling excess or cause-specific mortality, fixed assumptions about the two competing hazards can be made - those due to the disease and those due to other causes. This will lead to more explicit assumptions about the long-term all-cause marginal hazard function.

Relative survival models make the following assumptions:

- The hazard of death from other causes can be approximated for the population of interest through a population lifetable, which are typically assumed to be fixed and known rates with no associated uncertainty.
- That by isolating the excess mortality, the shape of the short-term hazard can be better captured, then other-cause mortality can be re-introduced through an external data source.
- That both disease-specific and other-cause mortality can be extrapolated separately, and successfully.

### 3.7.3. *Limitations*

Excess mortality models are subject to the following limitations:

- There is a need to specify a relevant external population in order to isolate the disease-specific mortality. That is, a population that is exchangeable in other-cause rates for the cohort of individuals with the disease of interest; it may be necessary to match on more factors, such as smoking and other lifestyle characteristics depending on the disease.
- The approaches are often used in large population-based studies and require the effect of age and other important determinants of each competing cause-specific hazard rate to be accounted for; this may be more challenging with a limited sample size in a trial setting.
- A choice of modelling approach with the relevant complexity to capture the shape of the excess mortality in the period of the trial is still necessary.
- Further, assumptions are still necessary about the long-term extrapolations of the disease-specific and other-cause mortality; that is, the long-term hazard functions for each cause of interest must still be defined to extrapolate appropriately.

### 3.7.4. *Further issues*

Consideration should be given as to whether a cause-specific modelling approach may be more applicable in a trial setting where closer attention is perhaps given to correctly assigning the cause of death with relation to the disease of interest. Excess mortality approaches are preferred in settings of population-based data, where information on cause of death is less reliable for the purpose of ascertaining if the death was due or not to the disease of interest, particularly for elderly patients, who may well be excluded from the trial setting at the outset. Whether or not cause-specific death information is collected in clinical trials, the concept of modelling disease-specific mortality and other-cause mortality separately for the purpose of extrapolation to inform economic modelling may be worthy of further consideration. It should be noted that taking a cause-specific approach rather than an all-cause approach will decrease the number of events, impacting on uncertainty. Given that in most trials the disease of interest will likely dominate the short-term mortality, this is unlikely to have a large impact for the short-term fit.

## 3.8. ADOPTING A BAYESIAN APPROACH

The potential advantages of adopting a Bayesian approach to modelling survival data (and other aspects of a HTA) include the ability to flexibly model evidence from a variety of data sources, to formally incorporate expert/clinical subjective prior beliefs, and to capture all forms of uncertainty (both parameter and model/structural) and which can be propagated through to eventual outcomes of interest, for example mean survival or net (monetary) benefit [36]. The formal inclusion of subjective prior beliefs has been greatly facilitated by the development of elicitation software[47]. Whilst the use of Bayesian methods applied to survival data (including extrapolation) in HTA have been advocated (Spigelhalter et al, 2003)[48] and user-friendly software developed (Baio, 2020)[49], in practice there have been relatively few fully Bayesian (using subjective informative prior distributions) applications either in the published literature or NICE TAs[50-52]. However, a number of HTAs have considered the use of subjective beliefs and/or external information regarding relative treatment effects, including the “borrowing of strength” using “class effects”[50-52]. A number of approaches to extrapolating survival data using Bayesian methods which include external information have been proposed (Abrams et al, 2016[13]; Guyot et al, 2017[53]; Soikkeli et al, 2019[54]). These generally fall into two broad approaches; firstly using the flexibility of a Bayesian approach to model (through the likelihood function) the different sources of information (including data from the trial under consideration) and adopting vague prior distributions for resulting model parameters[53], secondly by using a power prior approach for the external data with hyper-parameters possibly based on subjective beliefs about the plausibility of the external data relative to the trial and/or target population[55], or thirdly by using external information and/or subjective beliefs to specify prior distributions for either model parameters[54, 56] or prior model probabilities within a Bayesian model averaging framework[13]. In the case of using subjective beliefs this could be achieved by eliciting information on quantities experts/clinicians can readily express beliefs about, for example the proportion surviving at specific time points, and using this information to derive an implied prior distribution for model parameters[57]. In terms of survival mechanisms (and thus data generation mechanisms), though in an ecology and not a HTA setting, the inclusion of subjective beliefs regarding the plausibility of competing causes of death has also been considered[58].

In HTA, economic evaluation is undertaken to inform resource allocation decisions for disease populations, but survival analysis typically focuses on models fit to trial (sample) data. Fitting models under a Bayesian framework represents one option for moving away from trial-oriented analyses. Given the potential advantages, but limited practical application and evaluation to-date, further research in this area would be useful.

# 4. METHODS: EXTRAPOLATION OF SURVIVAL FUNCTIONS AND INCORPORATING EXTERNAL INFORMATION

## 4.1. BACKGROUND

Given that many clinical trials have restricted follow-up, and many diseases will result in a substantial proportion of individuals estimated to still be alive at that point, some form of extrapolation is often required in order to assess the lifetime impact of an intervention. This is used to populate economic models. In the previous section, we have discussed various approaches to fitting more complex survival models to capture the hazard functions seen within the range of the trial data; but many of the approaches we have introduced only make sense when coupled with external data (excess mortality and cure models, for instance), and a number of the approaches are tailored towards extrapolation rather than simply fitting within the range of the data.

In the process of extrapolation when using a statistical model two key considerations need to be made around capturing the shape of the hazard/survival function. Firstly, a sufficiently complex modelling approach should be used to capture the shape of the hazard within the range of the follow-up data; this ensures accurate prediction of the absolute risks within the range of the data and allows the extrapolation to "start from the right place". Secondly, thought needs to be given to the likely shape of the long-term hazard; this is unlikely to follow a consistent shape extrapolated from the model fitted during the period of the trial – which tends to capture mortality due to the disease in the short term. Those who have survived until the end of the trial are continually ageing and are likely to be impacted by competing risks from other cause mortality, but may also have a reduced risk of dying from the disease under investigation - though this may not be the case if a treatment simply delays disease-specific mortality rather than prevents it. These considerations should be taken into account when extrapolating survival functions. The methods described in Section 2 can allow complex hazard shapes to be captured *within* the range of the follow-up data. However, a simple extrapolation from these more complex methods will not necessarily lead to a good long-term extrapolation for the survival function. Some of the complex approaches will apply even more restrictive and unrealistic shapes on the long-term hazard than a standard, simple parametric approach. Therefore, thought should be given separately to the likely long-term hazard shape; which may require the utilisation of external data and the likely consideration of the fact that individuals will age as follow-up is extended.

In health technology assessment it is common for survival models fitted to clinical trial data to be used to extrapolate far into the future, with no external information taken into account. Sometimes extrapolations are compared to external evidence to provide a form of validation. It is relatively rare for external information to actually be incorporated in the model fitting process, although there are examples of this in the literature[53]. It is not the purpose of this document to make detailed recommendations on incorporating external information into survival models, but it is impossible to ignore this topic when discussing survival models that are used to extrapolate into the long-term. In this section, we consider the survival models introduced in Section 2 in the context of extrapolation and external evidence.

## 4.2. INCORPORATING EXTERNAL DATA

A number of methods have been suggested for the incorporation of external information to guide the extrapolation survival functions[59]. One assumption that is possible is to allow the long-term survival experience to be fully governed by population mortality rates or mortality rates for comparable patients in a relevant disease register. This could be achieved by applying the constraints discussed in Section 2.6 for the time-to-cure. The general population rates may be inappropriate if the trial population remain at an excess risk of death due to the disease of interest at, and beyond, the point of extrapolation. Alternatively, a relative or additive comparison to external rates can be made in order to extrapolate an effect that may not be appropriately defined by the external data. Further assumptions could be that the external data is not fully reflective of the trial data and a relative or additive difference between the trial data and external data is required in order to correct for this difference.

Many would argue that using a relevant disease register to identify the external data would be preferable, as these represent patients diagnosed with a similar condition and are likely to be more representative of the trial population. However, registries will not include patients being treated with new treatments, and therefore whilst they may be useful for modelling long-term survival for trial control groups, their usefulness for modelling long-term survival for people in the experimental group of clinical trials needs careful consideration. How long will the effect of a treatment be sustained? And will this act in a proportional manner on the hazard function for the control arm up until that point?

General population mortality rates are readily accessible, even if they are only stratified by broad factors, such as age, sex and calendar time. Whilst it is highly unlikely to be appropriate that all patients that remain alive at the end of a clinical trial have mortality rates equal to those observed in the general population, such data can at least be used to ensure that extrapolations do not result in long-term hazards that are below those observed in the general population. This involves the extrapolation issue being set into the context of a competing risk problem; there is the mortality associated with the disease of interest and mortality associated with other causes, with the former likely to dominate in the short-term (within the range of the trial), and the latter likely to dominate in the long-term for those that survive. Isolating the cause-specific mortality from an all-cause survival model can be achieved by partitioning the all-cause mortality into component parts, using general population mortality rates as the mortality experience due to other causes and relating any excess mortality above this to the disease of interest (see Section 2.7 on excess mortality/relative survival models). This relies on the assumption that the background general population mortality rates are a suitable match for the mortality experience of patients who do not die from their disease; this may not be the case for certain conditions where the likelihood of comorbidities will mean that general population mortality rates may be too low. Should instead cause of death information be available then obtaining the cause-specific mortality using standard parametric modelling is also a viable alternative. In both settings, careful consideration should be given to also modelling covariate effects to arrive at an appropriate marginal estimate of all-cause survival.

This approach of treating the cause-specific hazard separately makes sense when the other cause mortality is likely to dominate in the long-term; in these instances, extrapolating using external data is the logical approach. However, matching to registry or population mortality data to obtain the other cause mortality estimates relevant for patients relies on having access to the patient-level trial data; it is necessary to, at a minimum, match on age, sex and calendar time to appropriately reflect the risk in mortality, which will vary considerably.

In the following paragraphs, each of the approaches outlined in Section 2 are considered in the context of extrapolation. Consideration is given as to whether it would be appropriate to extrapolate from the fitted model to the trial data, and also how external data could be incorporated to guide the extrapolation in the long-term. In practice, the incorporation of external data for extrapolation has perhaps been constrained to the final section of a piecewise approach, or to couple with a cure model approach to estimate the other-cause hazard. However, external data can be coupled with other standard survival modelling approaches as outlined in the following subsections. The hazard function for other-cause mortality should always be considered when justifying the marginal hazard and survival functions used for extrapolation even if external data are not used in their estimation.

## 4.3. EXTRAPOLATION FOR COMPLEX SURVIVAL APPROACHES

Any model-based extrapolation from the methods that do not routinely include external data (Sections 3.3.1-3.3.4) are unlikely to appropriately account for changes in the hazard beyond the range of the observed data that are likely due to ageing effects increasing the long-term other cause mortality.

### 4.3.1. *Flexible parametric models*

FPMs can be used as an approach to capture complex hazard shapes within the range of the data. Beyond the final knots, the log(-log) transformation FPM described in Section 2.2 is similar to a Weibull model in that the log cumulative hazard is a linear function of log time. In principle, this can be used to extrapolate to a life-time horizon. This approach shares some of the limitations associated with using any standard parametric model applied to the range of the data and then used to extrapolate, although the approach differs slightly in that the long-term hazard is dictated by the latest mortality. Effectively, there is little to guide where the extrapolation should go and there is no option to prevent an unrealistic extrapolation; particularly for long-term estimates where the effect of age will likely begin to have a strong impact on increasing the hazard function.

It is possible to couple the excess mortality modelling approach (utilising external data) using a flexible parametric modelling framework (see Section 2.7). Furthermore, as for a standard parametric model, it is possible to make assumptions about a continuing relative or additive effect when comparing the clinical trial data to external data sources.

### 4.3.2. *Mixture models*

Again, within the observed range of the data, a mixture modelling approach may very well be suited to capturing more complex hazard functions, allowing a closer approximation to the observed hazard shape. However, extrapolations from these models may end up being unrealistic without the consideration of some external data to govern how the mortality rates should look in the longer term. If there is information on different clinically meaningful subgroups of patients, then this might be used to better inform the long-term extrapolations.  Typically, extrapolation from a complex fitted model is considered to be high risk in that extrapolated functions may be unstable. The mixture components are defined in order to provide the best fit to the observed data, but no constraints are in place to ensure that when extrapolating for higher values of follow-up time that the estimates will be in any way reasonable. External data should be used to at least govern whether reasonable extrapolations are obtained if one chooses to use the fitted model to estimate the lifetime horizon estimates.

### 4.3.3. *Landmark*

Landmark models seek to stratify the patients still alive at the landmark times into groups that are likely to be more similar in survival experience. Typically, standard parametric models are then fitted beyond the landmark point in order to provide the extrapolation. Therefore, the same limitations associated with extrapolating with standard survival models apply. In fact, selecting a group with good prognosis as in this case, may exacerbate the issues in certain circumstances, as very long-term survival may be predicted for this group, with very low hazards potentially still being predicted as the remaining patients age. External data should be used to at least govern whether reasonable extrapolations are obtained when extrapolating far into the future. As for any other survival model, external information could be incorporated into the survival modelling process to avoid clearly implausible extrapolations.

### 4.3.4. *Piecewise*

Partitioning the survival experience into sections and fitting separate models to each still relies on a final extrapolation from a parametric model. Again, these are unlikely to appropriately account for changes in the hazard beyond the range of the observed data, which are likely in the long-term because of the effect of ageing. A further consideration is the increasing influence of random variation when fitting survival models relying on fewer events in the interval. A standard extrapolation of the parametric models fitted to the final period coupled with the data for the previous periods will give an extrapolated experience for the entire treatment arm. Again, external information could be incorporated into the survival modelling process to avoid clearly implausible extrapolations.

### 4.3.5. *Cure*

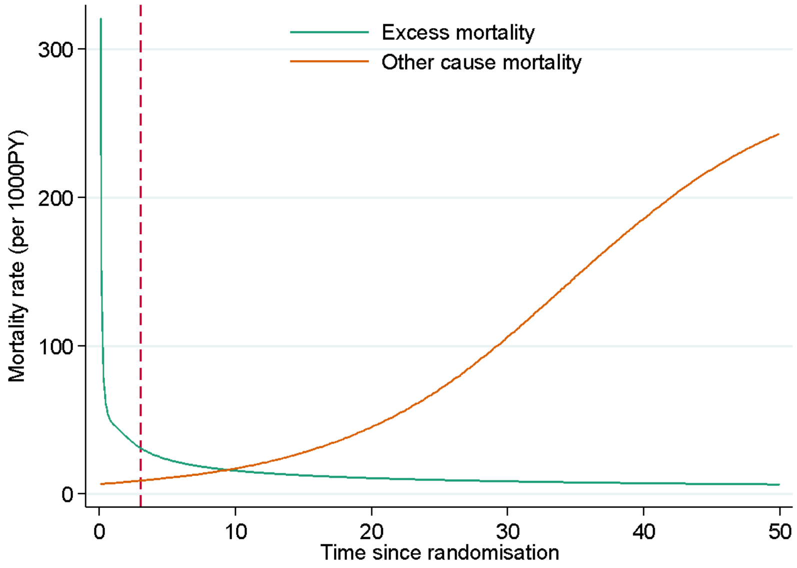
External information is crucial for cure models – other cause mortality must be incorporated, otherwise a proportion of patients will be predicted to never die. A cure model could be fitted to cause-specific data, but again extrapolation without other cause mortality information would be nonsensical, as it is essential to include information on other cause mortality to provide estimates of all-cause mortality. The cure model in the setting of excess mortality or cause-specific mortality is a useful approach for setting a timepoint at which the external data completely governs the long-term mortality experience. Within a flexible parametric modelling environment, it is possible to set the point at which cure is dictated to happen by the placement of the last knot of the spline function (in the context of flexible parametric cure models; extra constraints are placed to force cure beyond the final boundary knot). This could be a point beyond the time range of the trial, forcing the cause-specific mortality to plateau at that future point, and then letting other cause mortality take over, perhaps based upon population mortality rates. This offers a contrasting approach to extrapolating the excess mortality (see below).

### 4.3.6. *Excess mortality / cause-specific mortality (relative survival)*

In a competing risks setting, it is possible and of interest to consider cause-specific survival endpoints within the range of the trial. However, to extrapolate one must then first consider the all-cause survival function by re-incorporating other cause mortality; this could be from population mortality rates. This approach of extrapolating isolated functions and combining has been applied to extrapolate all-cause survival functions in a population-based cancer setting[10]; for the cause-specific model, one can either assume cure, or model the long-term cause-specific survival which will have typically plateaued and be dominated by the other cause hazard in the long-term. This approach to extrapolation allows a hazard function with a complex shape in the sense that the long-term external data can capture the increases associated with population ageing (see Figure 8).

In this approach, external data is already built into the initial modelling process and is therefore available for use in the extrapolation. Assumptions are typically needed about the long-term background mortality – this can be modelled taking into account the effect of calendar year, or it could be assumed that recent rates are reflective of what will happen in future.

**Figure 8: Illustrative example of isolated competing hazard functions over time**



## 4.4. CONSIDERATIONS FOR ALL EXTERNAL DATA APPROACHES

*Are our external data appropriate?*

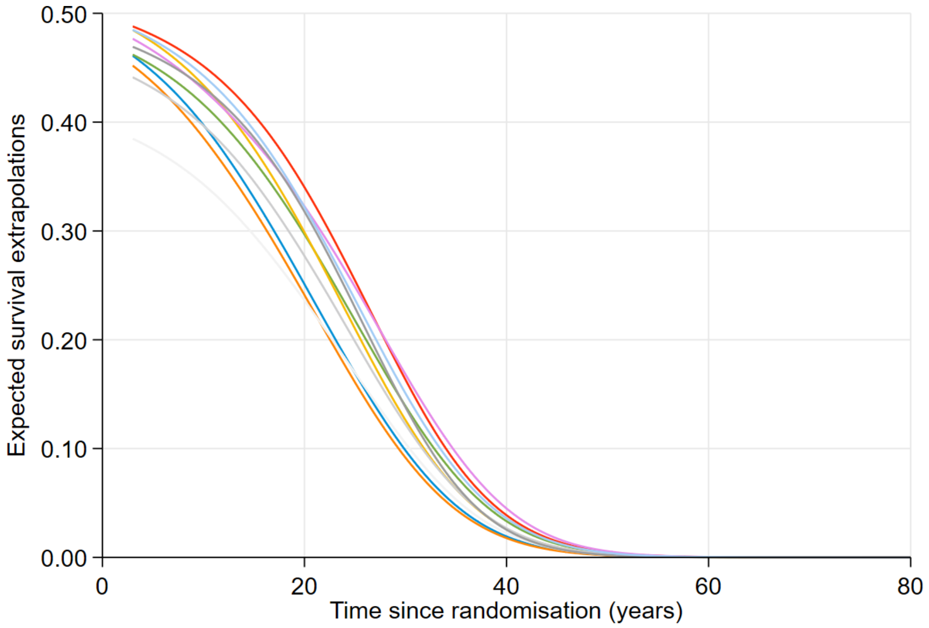
In all approaches we need to assume that we have fully captured the long-term mortality experience for the disease population using external data; or a function of an external data resource. Relevant considerations when using external data sources are:

- How well the external data source would truly reflect the survival experience should we have had full long-term data on all patients.
- The profile of the patients remaining at the end of follow-up.
- The length of time needed to reach the point at which all patients would have died.
- The proportion of patients for which we need to extrapolate.

*Small sample size*

One factor to consider in a clinical trial setting is the relatively small sample size, particularly towards the end of follow-up when fewer patients are still at risk. This may be particularly important when considering joining different survival models together, for example by appending a long-term survival function onto a curve fitted to survival data (such as in a piecewise approach). This is illustrated in Figure 9. We simulated a trial arm with 100 observations from a Weibull distribution, where survival data were available for 3 years. Beyond 3 years, extrapolations were based solely on background population mortality rates. Figure 9 illustrates post-3 year survival predictions for 10 runs of the simulated trial, illustrating that, due to random variation, the start-point of the extrapolation varies considerably, leading to considerably different long-term survival functions even when the survival data are from the same underlying model and the same external data are used to extrapolate. In this example, mean life years varied from 9.3 to 11.3 years across the 10 simulations. The variation stems from the fact that the sample size is 100 and also the variation in the age distribution at the end point of 3 years, which has an impact on the shape of the extrapolated curve.

**Figure 9: Illustration of varying survival extrapolations from 10 simulated trial arms with the same underlying disease-specific hazard and using the same external data**



*Does our trial data (in terms of patient characteristics) reflect the population target?*

A consideration for all clinical trial data, but also of particular focus when extrapolating the survival function, is whether the study population reflects the population for which the intervention will be potentially used. The relative effect may possibly remain unchanged across different populations, but the absolute survival functions and mean survival will vary if a different population is considered in its place. This influences both how we may wish to use clinical trial data to inform survival models, and also which external information to use when extrapolating.

*Are patient-level data for the trial available for the extrapolation?*

Patient-level data from relevant clinical trials are highly desirable when incorporating external data as this typically requires matching on characteristics such as age and sex, at least.

*Are patient-level data for the external data available for the extrapolation?*

Should the patient-level data be unavailable for relevant external data, then 1:1 matching between trial patients and patients in external data is not possible. An alternative may be to obtain the marginal hazard data for patients in the external dataset who have similar characteristics (such as age and sex) to patients included in the clinical trial. This could then be used beyond the end of the trial as the mortality experience for the extrapolation, or at the very least be used as a check to examine the plausibility of the extrapolation from another approach.

# 5. SIMULATION STUDIES

## 5.1. INTRODUCTION

A simulation study was conducted to assess the performance of various survival models to estimate the restricted mean survival time at the end of a trial follow-up and the overall mean survival based on extrapolation, in a range of biologically plausible scenarios based on a number of previous TAs (TA268[41], TA391[38], TA421[29], TA463[10], TA478[9], TA483[34], TA498[7], TA517[8] & TA519[39]). The aim was to evaluate extrapolation in a single treatment arm rather than treatment effects because, fundamentally, it is crucial to extrapolate survival as accurately as possible when undertaking economic evaluation of treatments that affect survival. If survival models are unable to extrapolate accurately, they are inappropriate for economic modelling irrespective of what they predict with respect to the treatment effect. The simulation study is reported in line with recommendations made by Morris et al[60]. However, it should be noted, that the assumptions surrounding the treatment effect in the short, and long-term are a crucial driver of any metric for the difference in mean survival. We make further recommendations about the treatment effect in Section 6.

## 5.2. METHODS

### 5.2.1. *Data generating mechanisms*

The simulation study considered both small ($n = 100$) and medium ($n = 500$) sized trials, and trials with low and medium survival rates with a follow-up of 3-years. In addition, four scenarios were considered based on the true distribution of the disease-specific survival function; and each scenario further incorporated small and large amounts of unmeasured heterogeneity in individual survival functions[61]. This gave 32 data-generating scenarios in total. All scenarios incorporated background mortality rates to represent age-related other causes of death.

Other-cause mortality was simulated from a Gompertz distribution to represent background mortality unrelated to the disease of interest. The ages of trial participants were simulated from a Normal distribution with a mean age of 60 and standard deviation 6. Then conditional on survival to current age $a$, the survival function for other-causes was:

*Equation 15*

$$S_{other}(t|a) = P(T \geq t + a \mid T \geq a) = S_{other}\,(t + a)/S_{other}\,(a)$$

Where,

*Equation 16*

$$S_{other}(t) = \exp\left(\lambda\gamma^{-1}(e^{\gamma t} - 1)\right)$$

and $\lambda = 0.000028, \gamma = 0.0936$ are the shape and scale parameters, respectively. These shape and scale parameters were obtained from fitting a Gompertz distribution to English mortality rates in 2009 in females.

Disease-specific mortality was simulated from a two-component Weibull mixture distribution with disease-specific survival $S_d(t) = S_{d0}(t)^{\exp(Z\beta)}$ where

*Equation 17*

$$S_{d0}(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1 - p)\exp(-\lambda_2 t^{\gamma_2})$$

and where $Z \sim N(0,1)$ is an unknown heterogeneity (frailty) term. This distribution allowed consideration of scenarios where the disease-specific survival function was Weibull ($p = 1$), a two-component Weibull mixture ($0 < p < 1$) or where the survival function corresponded to a cure model ($p \neq 0$and $\lambda_1 = 0$) with cure fraction $p$. The disease-specific survival scenarios considered were as follows.

*Scenario 1: Survival times simulated from a Weibull distribution with decreasing hazard*

| | Low survival | Medium survival |
|---|---|---|
| Low heterogeneity | $\lambda_1 = 0.55, \gamma_1 = 0.9, \beta = 0.5,$ $p = 1$ | $\lambda_1 = 0.25, \gamma_1 = 0.9, \beta = 0.5,$ $p = 1$ |
| High heterogeneity | $\lambda_1 = 0.55, \gamma_1 = 0.9, \beta = 2.0,$ $p = 1$ | $\lambda_1 = 0.25, \gamma_1 = 0.9, \beta = 2.0,$ $p = 1$ |

*Scenario 2: Survival times simulated from a Weibull distribution with increasing hazard*

| | Low survival | Medium survival |
|---|---|---|
| Low heterogeneity | $\lambda_1 = 0.10, \gamma_1 = 2.5, \beta = 0.5,$<br>$p = 1$ | $\lambda_1 = 0.04, \gamma_1 = 2.5, \beta = 0.5,$<br>$p = 1$ |
| High heterogeneity | $\lambda_1 = 0.10, \gamma_1 = 2.5, \beta = 2.0,$<br>$p = 1$ | $\lambda_1 = 0.04, \gamma_1 = 2.5, \beta = 2.0,$<br>$p = 1$ |

*Scenario 3: Survival times simulated from a Weibull mixture distribution with a high initial hazard that decreases and stabilises*

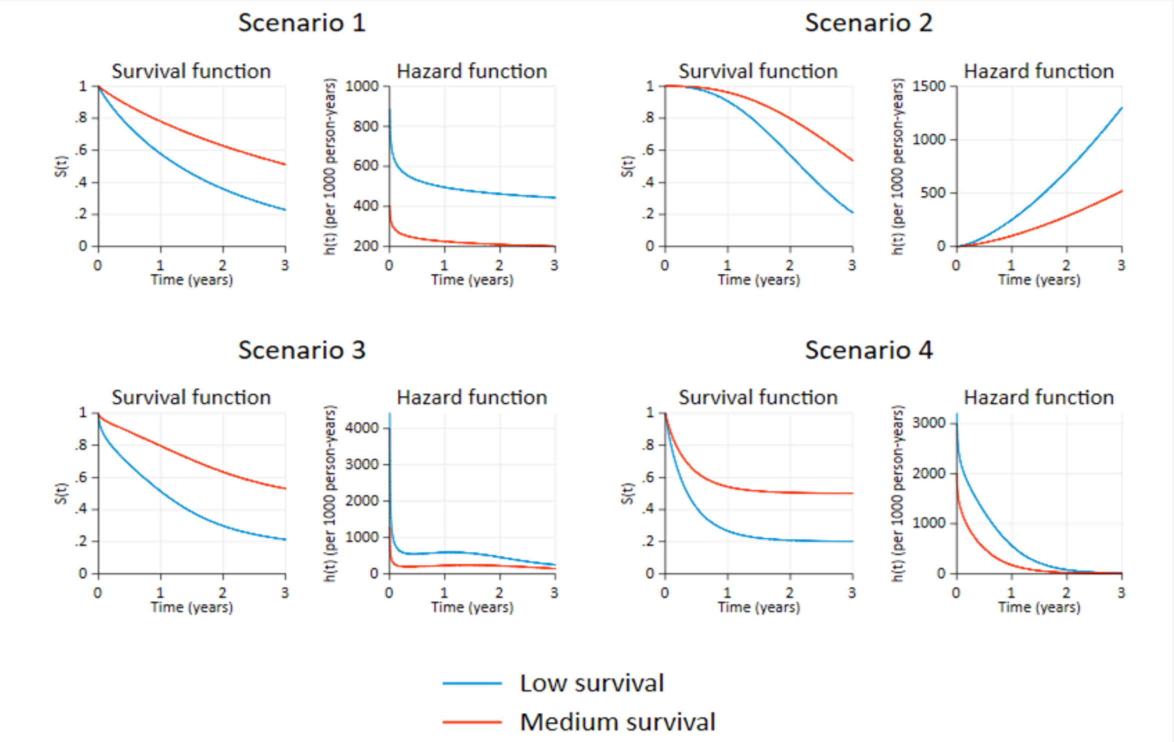
| | Low survival | Medium survival |
|---|---|---|
| Low heterogeneity | $\lambda_1 = 0.60, \gamma_1 = 1.8, \beta = 0.5,$<br>$p = 0.3$<br>$\lambda_2 = 0.70, \gamma_2 = 0.5$ | $\lambda_1 = 0.30, \gamma_1 = 1.8, \beta = 0.5,$<br>$p = 0.3$<br>$\lambda_2 = 0.20, \gamma_2 = 0.5$ |
| High heterogeneity | $\lambda_1 = 0.60, \gamma_1 = 1.8, \beta = 2.0,$<br>$p = 0.3$<br>$\lambda_2 = 0.70, \gamma_2 = 0.5$ | $\lambda_1 = 0.30, \gamma_1 = 1.8, \beta = 2.0,$<br>$p = 0.3$<br>$\lambda_2 = 0.20, \gamma_2 = 0.5$ |

*Scenario 4: Survival times simulated from a cure model with a Weibull distribution for the uncured*

| | Low survival | Medium survival |
|---|---|---|
| Low heterogeneity | $\lambda_2 = 2.5, \gamma_2 = 0.9, \beta = 0.5,$<br>$p = 0.2$ | $\lambda_2 = 2.5, \gamma_2 = 0.9, \beta = 0.5,$<br>$p = 0.5$ |
| High heterogeneity | $\lambda_2 = 2.5, \gamma_2 = 0.9, \beta = 2.0,$<br>$p = 0.2$ | $\lambda_2 = 2.5, \gamma_2 = 0.9, \beta = 2.0,$<br>$p = 0.5$ |

The disease-specific survival and hazard functions for Scenarios 1-4 are shown in Figure 10 for individuals with frailty $Z = 0$ for the 3-year follow-up period.

**Figure 10: Disease specific survival and hazard functions, Scenarios 1-4**



From the disease-specific and other-cause survival functions, cause-specific survival times were generated using the survsim function in Stata[62]. Overall survival was calculated as $T = \min(T_d, T_{other})$ and both random censoring and administrative censoring were applied. Random censoring was generated from an exponential distribution with rate 0.1 with administrative censoring applied to all survivors at 3-years. In total, 1000 datasets were simulated for each of the 32 data-generating scenarios.

### 5.2.2. *Estimands*

The estimands of interest were the restricted mean survival time (RMST) at 3-years (corresponding to the end of trial follow-up) and overall mean survival. From the data-generating mechanism these quantities cannot be evaluated using a closed form solution. Instead, for each data-generating scenario, one very large dataset was generated ($N = 10^7$) and numerical methods were used to evaluate the area under the survival function up to 3-years (giving the true value of the RMST at 3-years) and up to 80 years (giving the true mean survival time).

The true all-cause survival functions and hazard functions are shown in Figure 11-14.

**Figure 11A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 1**

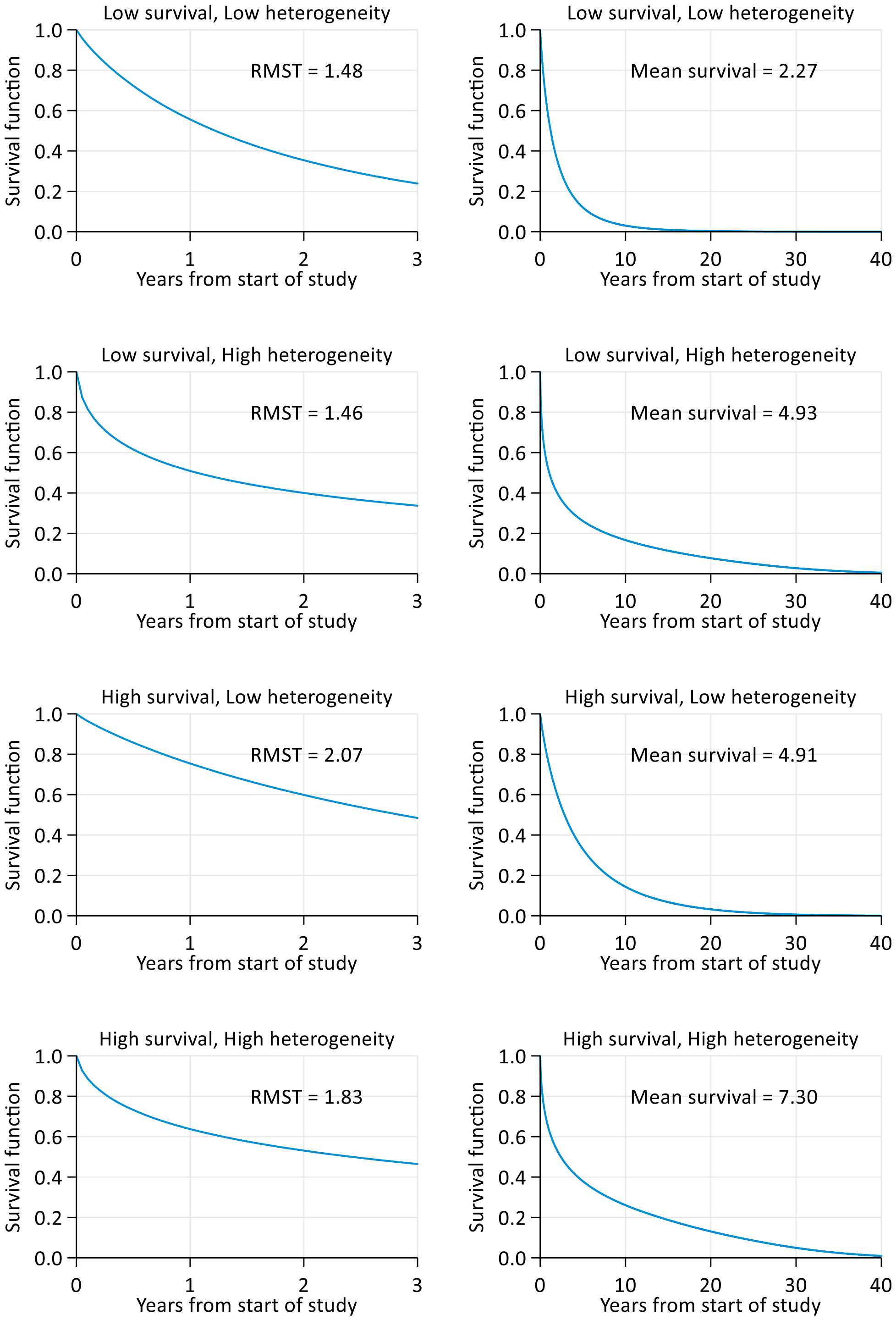


**Figure 11B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 1**

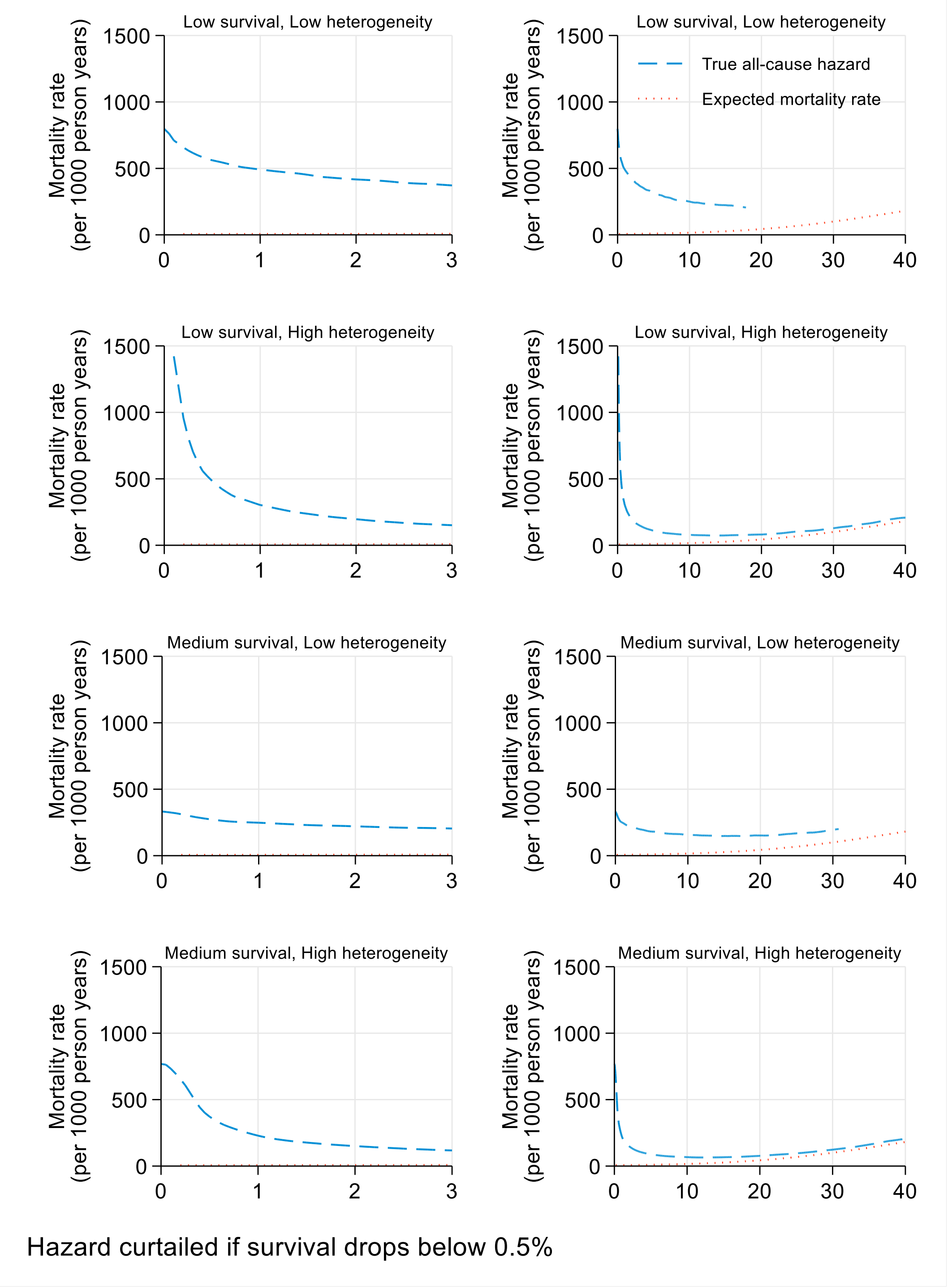


**Figure 12A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 2**

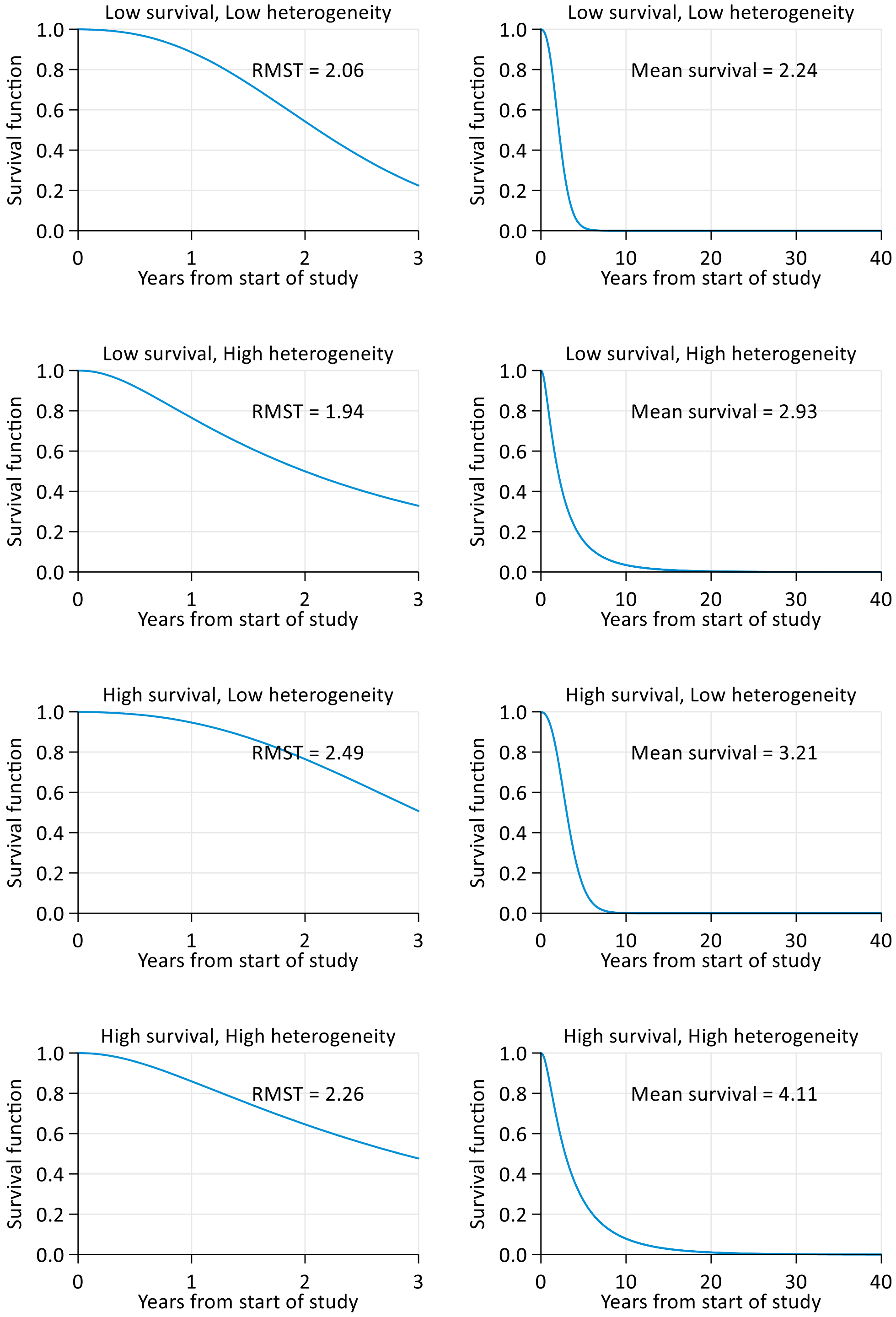
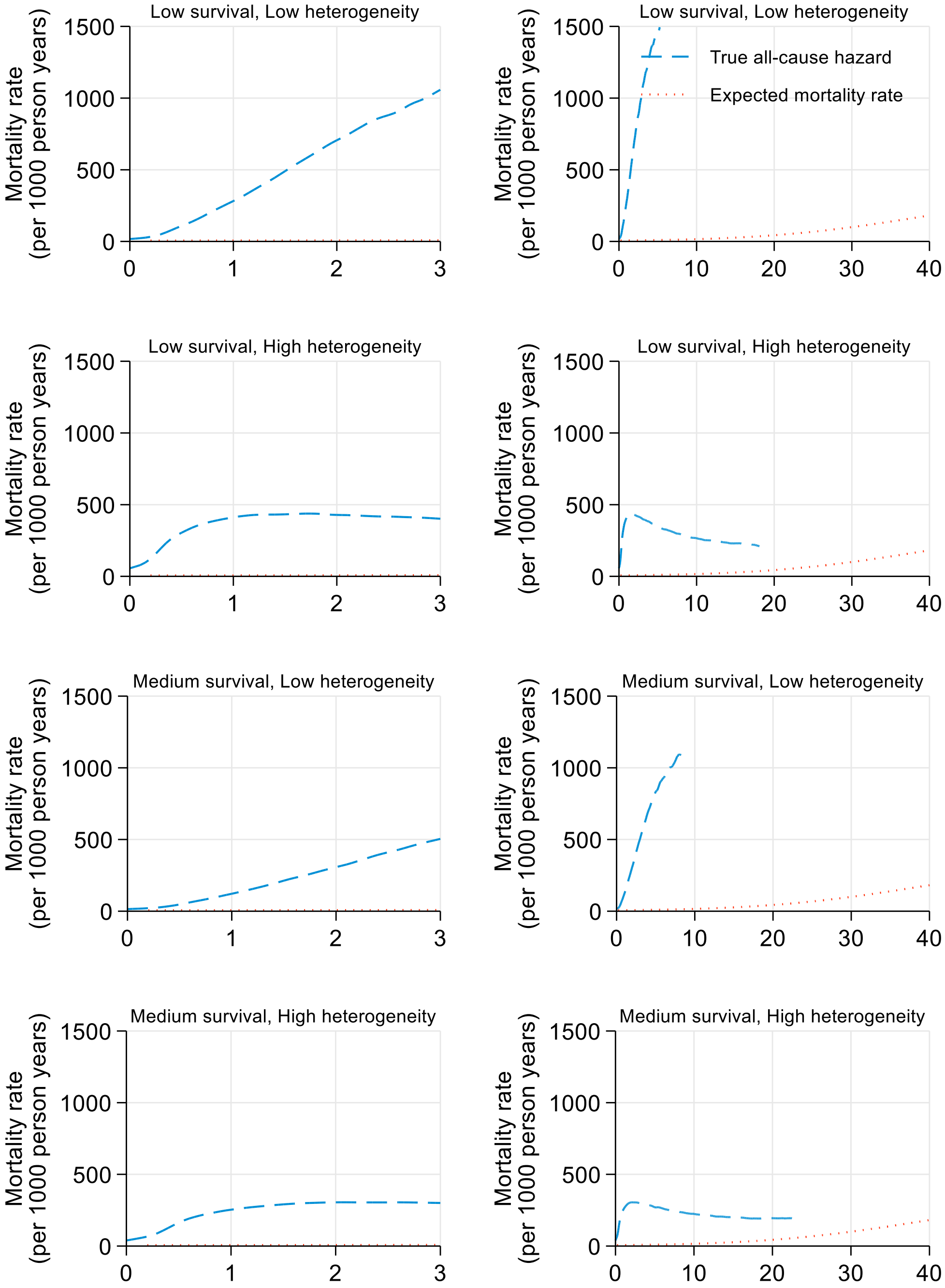


**Figure 12B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 2**



Hazard curtailed if survival drops below 0.5%

**Figure 13A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 3**

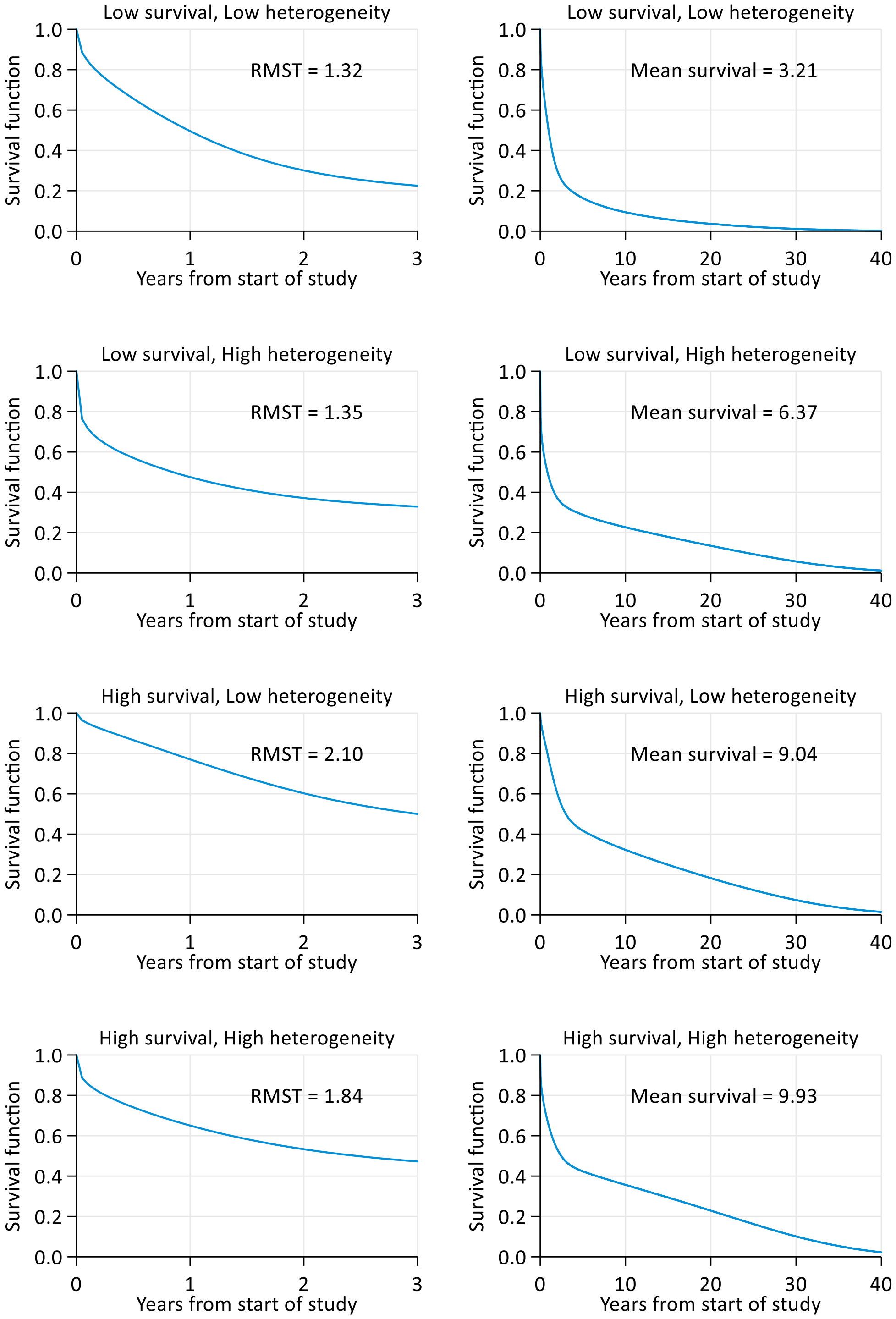
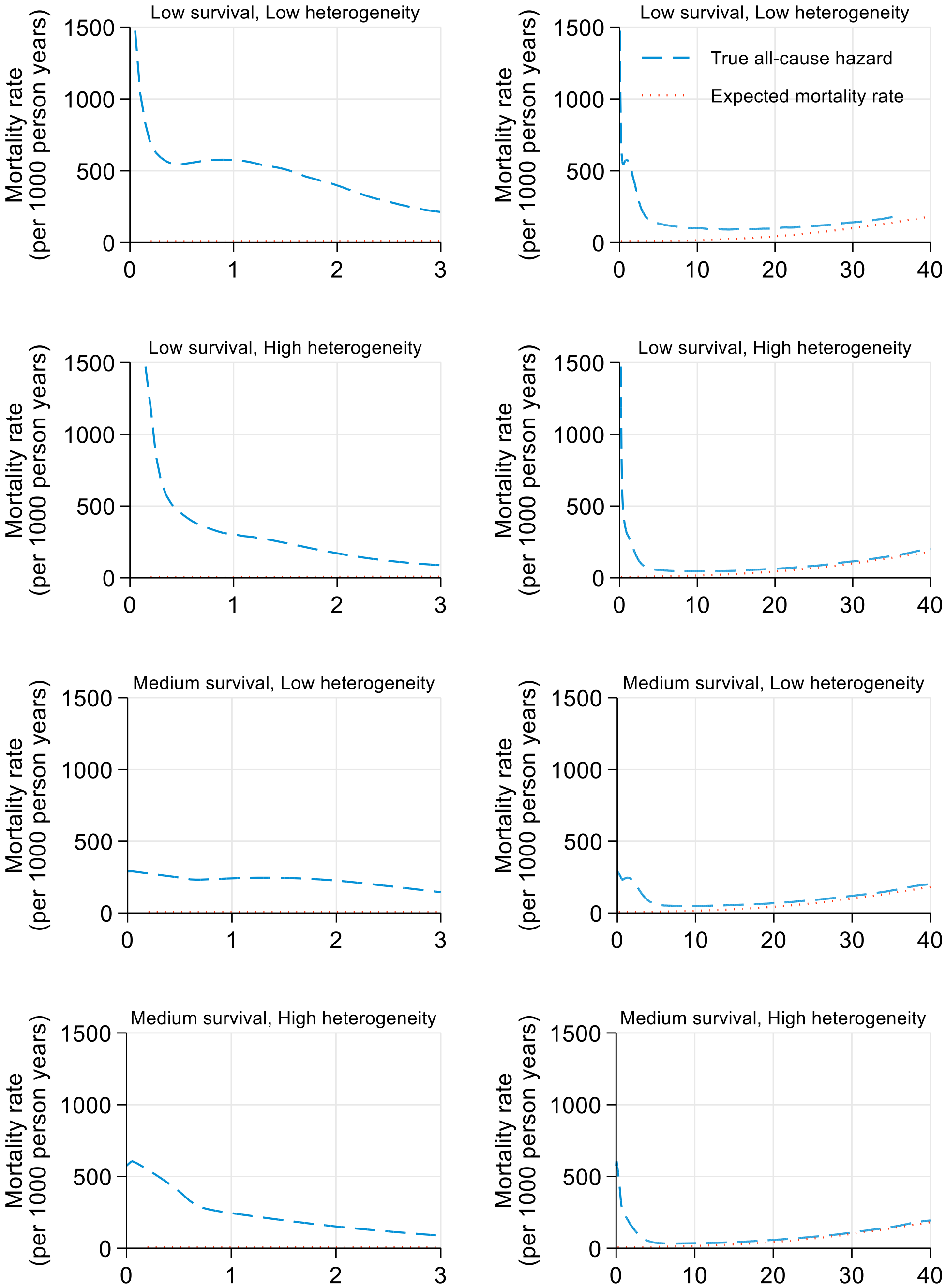


**Figure 13B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 3**



Hazard curtailed if survival drops below 0.5%

**Figure 14A: True survival functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 4**

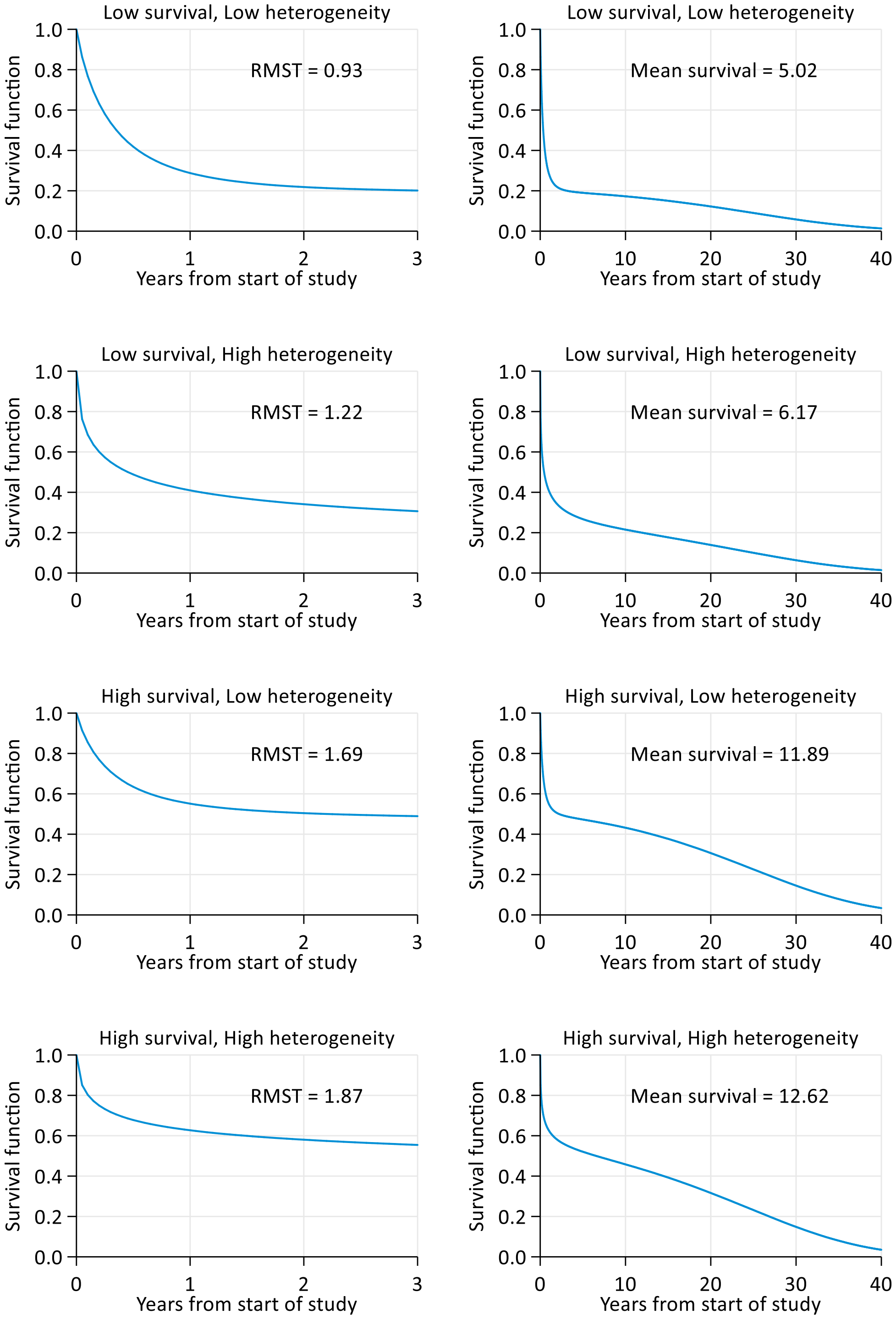
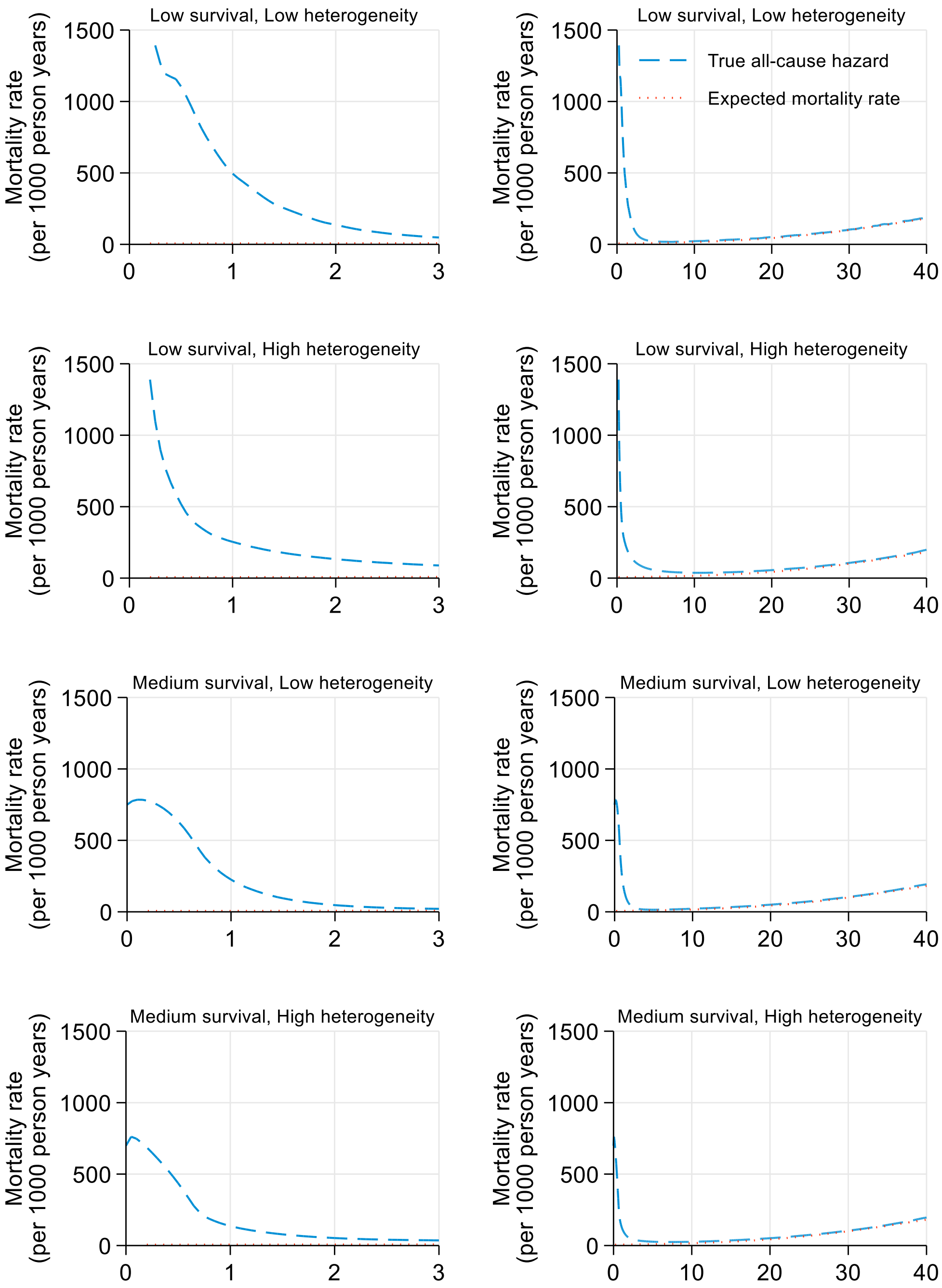


**Figure 14B: True hazard functions at 3 years (left hand side) and lifetime (right hand side) for Scenario 4**



Hazard curtailed if survival drops below 0.5%

### 5.2.3. *Survival models*

The models investigated were fitted to all-cause mortality from the simulated datasets. Simple models fitted directly to the data (without inclusion of background mortality) included a suite of parametric survival models (Exponential, Weibull, Log-Normal, Log

Logistic, Gompertz, Generalised Gamma) and Flexible Parametric Models (FPMs) with degrees of freedom for the baseline hazard spline function ranging from 1 to 5. In addition, a strategy of selecting the best fitting of these 5 FPMs using the AIC was also assessed. More complex models that included background mortality rates were also investigated and included relative survival FPMs (degrees of freedom ranging from 1 to 5) and a cure model (with either a Weibull baseline hazard or a FPM with 3 degrees of freedom). A mixture model was also considered but was found to have poor convergence (in some scenarios over 50% of the models failed to converge). It was therefore decided to exclude this model in the main simulation result comparisons. Landmark models were not included because we did not simulate response categories and would have had to also add consideration of the timepoint for response, which would have added an extra layer of complexity to the simulation process and increased the number of data generating mechanisms required to specifically evaluate this one method. Piecewise models were not included for practical reasons, due to the requirement for selection of piecewise time-points – to an extent, these are explored in Section 5.

All models were fitted in Stata using the functions streg (for parametric models), stpm2 (for FPMs, including cure and relative survival FPMs), strsmix (for a Weibull cure model) and stmix (for the mixture models). For all models, estimates of the RMST at 3-years and mean survival based on extrapolations from each of the models were obtained. For the FPMs this was performed using the stpm2 post-estimation predict command. For the standard parametric models, the cure models and the mixture models, estimates of the RMST at 3-years and the mean survival were obtained using the integ function on a fine grid of (500-1000) predicted survival probabilities.

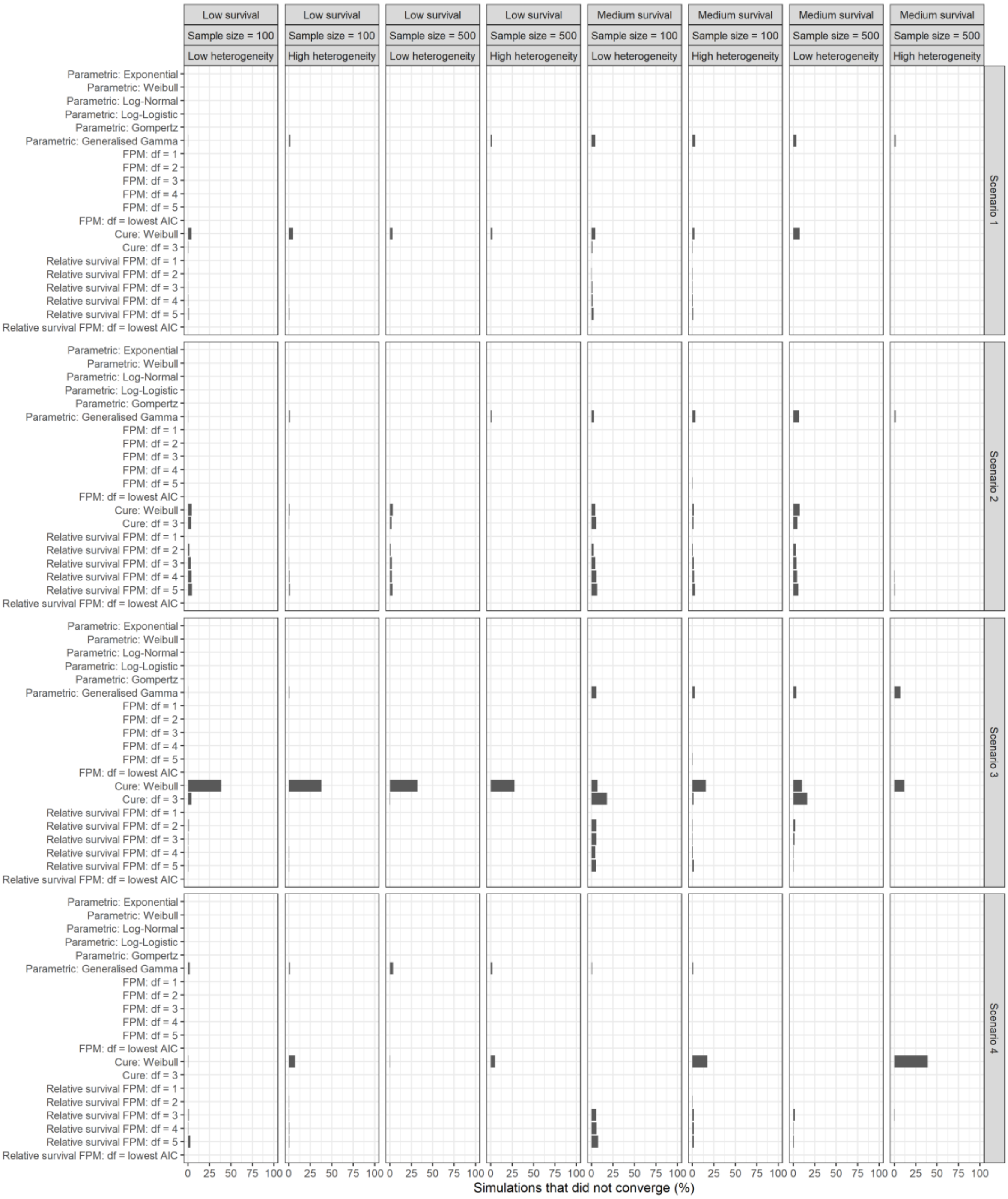
### 5.2.4. *Performance measures*

Bias of the estimates is of primary interest. Biases are shown in the form of forest plots with Monte Carlo standard errors. More detailed scatter plots of the biases are given in the appendices along with empirical standard errors in Appendix A4.

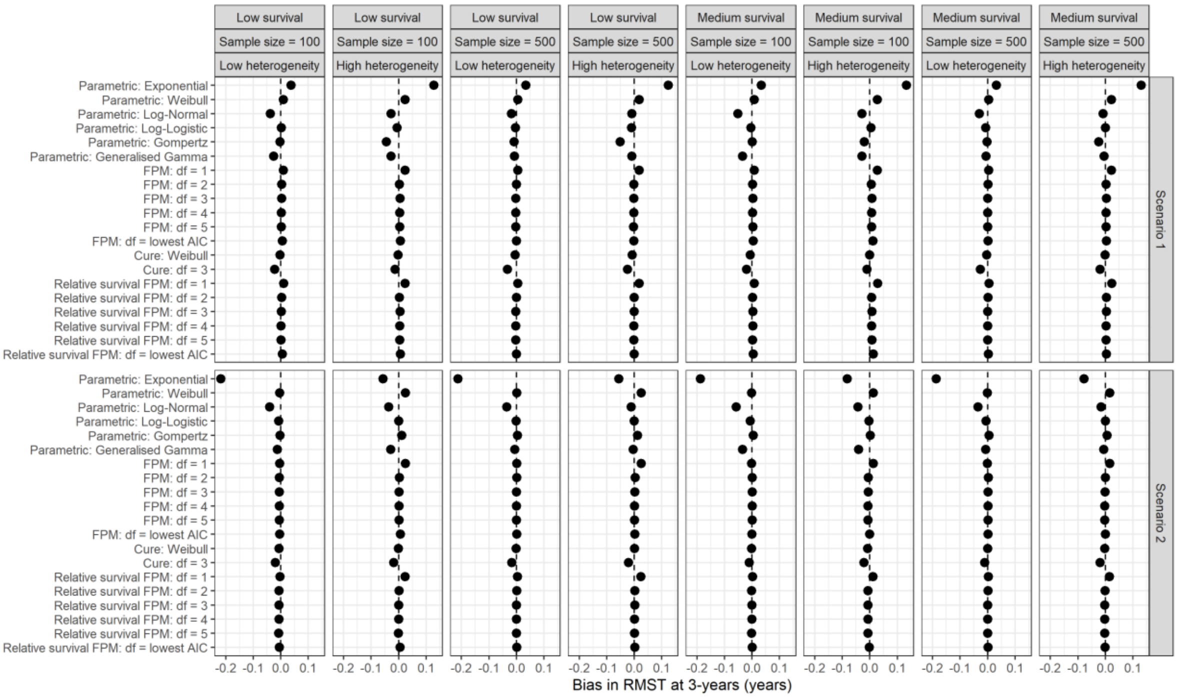
## 5.3. RESULTS

The appendix shows detailed results for all scenarios. These are summarised in Figures 15-19. We then interpret the results for each modelling approach.

## Figure 15: Convergence – percentage of simulations that failed to converge



## Figure 16: Bias in restricted mean survival time at 3-years, Scenarios 1 and 2



## Figure 17: Bias in restricted mean survival time at 3-years, Scenarios 3 and 4

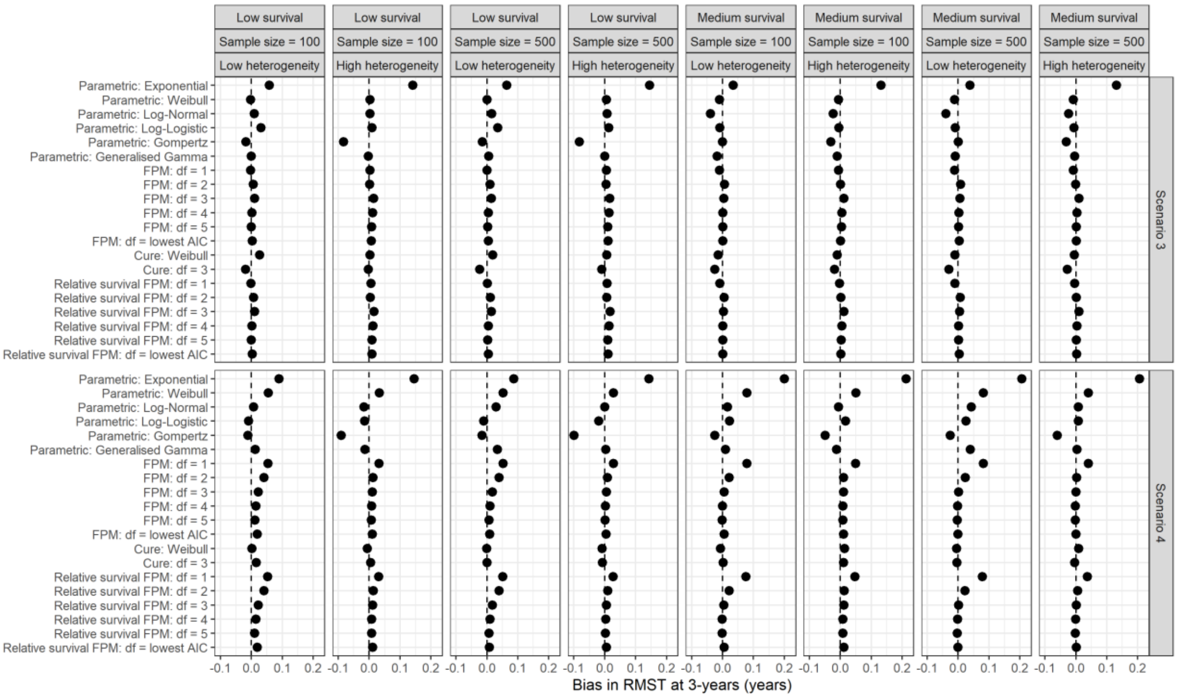


**Figure 18: Bias in mean overall survival times (axis restricted to a maximum of +/- 5 years, Scenarios 1 and 2**

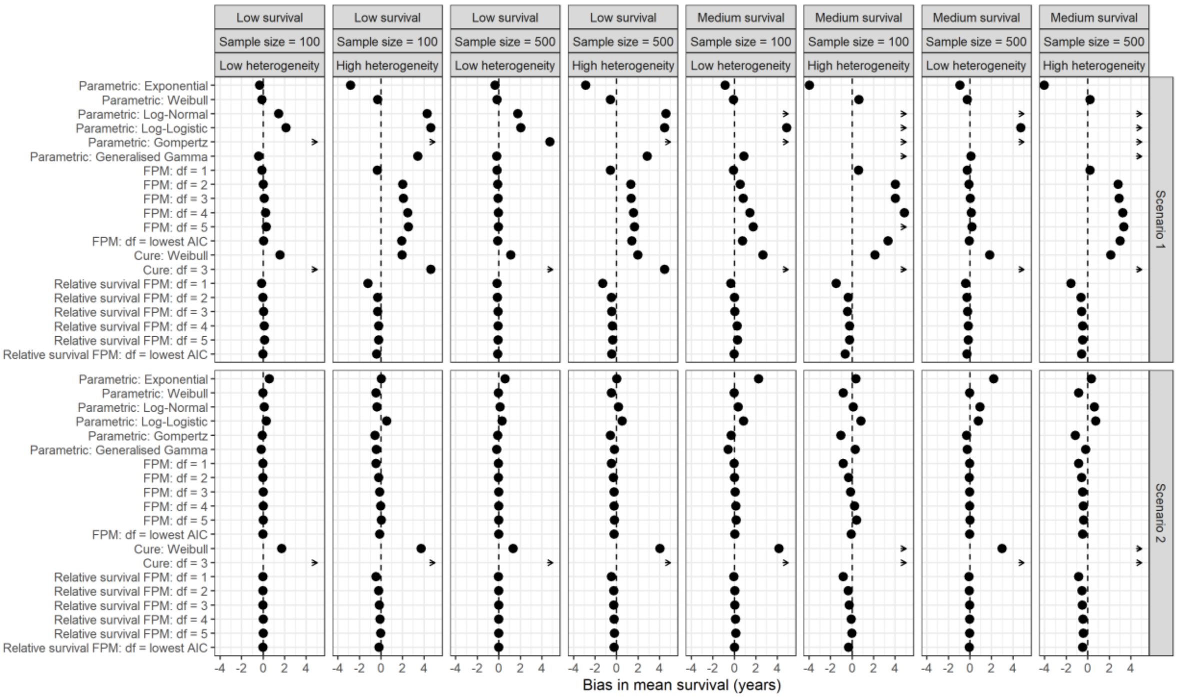
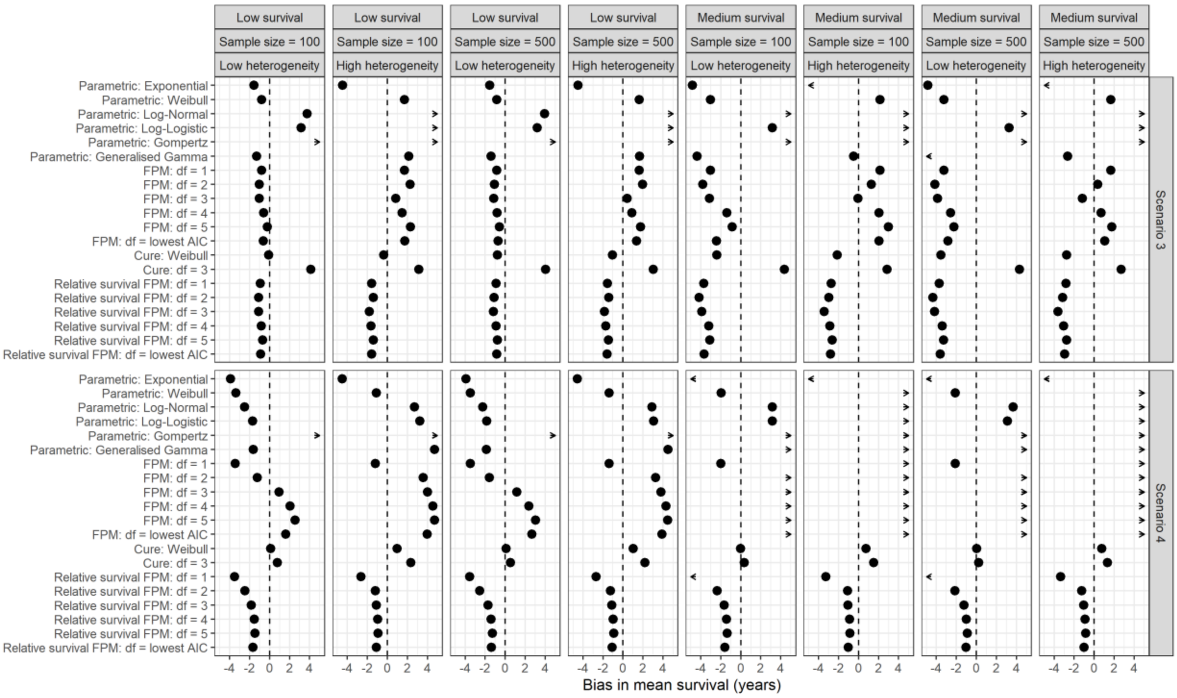


**Figure 19: Bias in mean overall survival times (axis restricted to a maximum of +/- 5 years, Scenarios 3 and 4**



### 5.3.1. *Standard Models*

Standard parametric models consistently produced low bias when estimating RMST (limited to the end of the simulated trial follow-up), except for the exponential distribution and, in some scenarios, the Gompertz distribution. Hence, models that do not well represent the underlying survival distribution are likely to result in only small bias for RMST. However, bias can become substantial when extrapolating to a life-time time horizon. Bias was greater with larger unobserved heterogeneity and when survival was relatively long (i.e. medium survival rather than low). This was true across Scenarios 1-4, irrespective of the distribution used to generate survival times, although bias was particularly high in Scenarios 1, 3 and 4, where survival followed a Weibull distribution with a decreasing hazard, a mixture Weibull and a Weibull distribution with a cure fraction respectively. Bias was lower in Scenario 2, where survival followed a Weibull distribution with an increasing hazard, whereby mean survival times were generally low. Across Scenarios 1, 3 and 4, the log-logistic and log normal models frequently over-estimated mean survival, except in Scenario 4 when there was small unobserved heterogeneity. Hence, even in scenarios with a cure fraction, these models – which are known to produce survival functions with long tails – often over-estimated mean survival. The Weibull model generally produced lower bias in its estimates of mean survival than the other standard parametric models – although often bias was still appreciable. This may be due to the way we generated our data using Weibull-based data generating mechanisms. However, although we simulated from a Weibull distribution for the disease-specific hazard, none of the standard models fitted are the correct model because we included unobserved heterogeneity and other cause mortality.

### 5.3.2. *Flexible Parametric Models*

FPMs were associated with very low bias for RMST across all scenarios, provided the FPMs had at least 3 degrees of freedom. Bias remained relatively low for estimates of mean overall survival in Scenario 2 (where survival followed a Weibull distribution with an increasing hazard). However, in Scenario 1 FPMs only produced low bias when survival was short and unobserved heterogeneity was small – when survival was long or heterogeneity was large FPMs produced high levels of bias. It is important to note that care should be taken in interpreting Figures 18 and 19 – biases that appear small are sometimes associated with mean survival estimate biases of 0.5 to 1 year; these are often not negligible. In Scenario 3 bias for mean overall survival was relatively low when survival was relatively short and heterogeneity was small (though, again, these biases were often not negligible) – bias increased appreciably when survival was long. Across Scenario 4, where there was a cure fraction, FPMs did not generally estimate mean overall survival with any less bias than standard parametric models and bias was often very large.

### 5.3.3. *Flexible Parametric Models (incorporating background mortality)*

FPMs that were fit using a relative survival framework, thereby incorporating background mortality, performed similarly to FPMs fit without incorporating background mortality in Scenario 2 – in which all FPMs produced relatively low bias in estimating mean overall survival. In Scenario 1 (where survival followed a Weibull distribution with a decreasing hazard) FPMs that included background mortality performed appreciably better than FPMs that did not incorporate background mortality. In Scenario 3, where survival followed a mixture Weibull distribution, including background mortality in FPMs did not result in an appreciable reduction in bias – models were prone to negative bias, under-estimating mean survival substantially, especially when survival was long. However, results appeared more consistent in Scenario 3 when background mortality was included, in that each FPM led to negative bias, whereas when FPMs without background mortality were fit some produced negative bias and some produced positive bias in this Scenario. The issues driving these differences are further explored in Section 5.6. In Scenario 4, where a cure fraction was simulated, including background mortality within FPMs led to an appreciable reduction in bias compared to standard parametric models (which did not include background mortality) and compared to FPMs that did not include background mortality. Mean overall survival was consistently under-estimated by FPMs that included background mortality in these scenarios, but bias was consistently lower than that associated with any other model tested, with the exception of cure models. However, note that bias remained approximately 1-2 years, which is appreciable given the true mean survival time varied between 5 and 13 years in Scenario 4.

It should be noted that in this simulation study we assumed that we had access to appropriate expected mortality rates from background mortality data. Incorporating this data results in improved extrapolation when there are some long-term survivors, but appreciable bias remains. This is likely to be because we assumed we only had data on 3 years of follow up, during which time disease-specific mortality rates may not stabilise. In Section 5 we demonstrate how increased follow-up can lead to better extrapolation.

### 5.3.4. *Cure Models (incorporates background mortality)*

Cure models (that incorporated background mortality) resulted in slight bias in RMST in scenarios where cure was not a reasonable assumption. However, these models still fit reasonably well to the observed data and produced low bias. Cure models led to substantial bias in estimates of mean overall survival in scenarios where a cure was not reasonable – in Scenarios 1 and 2, where other methods estimated mean survival with relatively low bias, cure models often resulted in appreciable bias. In Scenario 4, where a cure fraction was simulated, bias associated with cure models was lower than other methods, particularly when heterogeneity was low. When heterogeneity was high, cure models produced similar bias to FPMs that included background mortality, though bias was in the opposite direction – that is, mean overall survival was usually over-estimated. In Section 5 we demonstrate how varying the point of cure can alter survival estimates associated with cure models.

## 5.4. SUMMARY & DISCUSSION

Any simulation study is limited to the scenarios it investigates. We have selected 32 different data generating mechanisms and simulated both disease-specific and other cause mortality and incorporated unobserved heterogeneity. We specifically targeted these data generating mechanisms to cover a broad range of biologically plausible scenarios. Unobserved heterogeneity is often ignored in simulation studies. We view this as unrealistic as in any applied setting one would always expect the underlying rates of death to vary between individuals.

We also generate data from specific distributions. For example, Scenario 1 is generated from a Weibull distribution (with frailty). However, it is the marginal survival function which is of interest. The shape of the marginal hazard function can be very different to the conditional hazard functions. None of the models fitted represented “true” models given the data generating process. However, we acknowledge that it would be possible to simulate from alternative distributions, which might result in different methods performing “best”. Hence, our simulation study should not be used to conclude that one method is superior to another. Instead, we seek to highlight that certain methods will exhibit bias in some situations.

The simulation clearly indicates that there will be no one method that can be universally applied in all circumstances to obtain unbiased estimates. Sometimes, what appears to be a reasonable method can result in severe bias. In general, standard parametric models that do not incorporate external information extrapolate poorly. FPMs can improve upon this, particularly when relevant external information is incorporated. However, serious bias can remain, particularly when disease-specific mortality rates have not stabilised during the follow-up period. Cure models can result in reasonable estimates of mean overall survival, but only when a cure assumption is reasonable – otherwise these models can result in high levels of bias in scenarios where other methods perform relatively well.

In appendix A.1 we provide details of where to download the simulation code, such that others can repeat and adapt our simulations.

# 6. EXAMPLES OF MODEL EXTRAPOLATION FOR CHOSEN SIMULATION SCENARIOS

## 6.1. INTRODUCTION

In this section we demonstrate a number of key issues using some examples using the simulated data. In a real analysis setting there is no way of knowing whether the extrapolations are correct. The previous chapter highlighted that certain methods could result in bias and the magnitude of the bias varies between different scenarios. As we are using simulated data we can investigate different types of analysis and try to understand why certain methods are likely to result in bias through a more in depth analysis of single simulated datasets.

## 6.2. CALCULATING MARGINAL EXPECTED SURVIVAL AND HAZARD

In this section, to aid interpretation in the Figures, we have added marginal expected survival using background mortality rates:

*Equation 18*

$$S^*(t) = \frac{1}{N}\sum_{i=1}^{N} S_i^*(t)$$

where $S_i^*(t)$ is the expected survival for the $i^{th}$ subject. This will give the expected survival in a disease free population. If population expected rates are used it will give the expected survival in a similar group to the study participants in the general population.

In addition, the marginal expected hazard can be estimated as a weighted average of expected hazard rates.

*Equation 19*

$$h^*(t) = \frac{\frac{1}{N}\sum_{i=1}^{N} S_i^*(t)h_i^*(t)}{\frac{1}{N}\sum_{i=1}^{N} S_i^*(t)}$$

where $h_i^*(t)$ is the expected hazard rate for the $i^{th}$ subject. This can serve as a useful reference as it is unlikely that the mortality rate in patients with the disease of interest will be lower than that in the general population. This is particularly useful as one can compare the extrapolated hazard rate for the study population with that seen in the general population. In nearly all cases one would not expect the mortality rates of diseased individuals to be lower than that seen in the general population.

When incorporating background mortality the marginal all-cause survival function is

Equation 20

$$S(t) = \frac{1}{N}\sum_{i=1}^{N} S_i^*(t)\widehat{R}_i(t)$$

where $\widehat{R}_i(t)$ is the model based estimate of the relative survival function for the $i^{th}$ subject.

The marginal extrapolated hazard for the study population is

*Equation 21*

$$h(t) = \frac{\frac{1}{N}\sum_{i=1}^{N} R_i(t)S_i^*(t)(\lambda_i(t) + h_i^*(t))}{\frac{1}{N}\sum_{i=1}^{N} S_i^*(t)R_i(t)}$$

where $\widehat{\lambda}_i(t)$ is the model based estimate of the excess mortality rate  for the $i^{th}$ subject.

### 6.3. Assessing model fit within the range of the data does not guarantee good extrapolation

Within the range of the data there may be a number of possible survival models that give a good fit to the data. However, care should be taken when using model fit criteria to choose an appropriate model for extrapolation. This is illustrated in Figures 20 and 21, which show a number of survival models fitted to a single simulated dataset from Scenario 1 (where the true disease-specific hazard is Weibull with unobserved frailty and other-cause mortality increases with age).

Figure 20 plots (a) the fitted survival functions and (b) hazard functions up to the end of trial follow-up at 3-years. Within the range of the data there is good agreement between most of the fitted models, the non-parametric Kaplan-Meier curve and the true survival function. The model with the lowest AIC is the Log-Normal model and the model with the highest is the Weibull model.

All hazard functions are decreasing at the end of follow-up. This should cause alarm as these fitted functions will continue to decrease beyond the range of follow-up and given that those still alive will be ageing, the true hazard would be expected to start increasing at some point. This is exactly what is seen in Figure 21, where the survival and hazard functions are extrapolated up to 40-years. The expected mortality (an age-sex matched rate based on population mortality rates) is seen to increase over the follow-up and therefore in the trial population the all-cause true hazard has a turning point at around 12 years of follow-up.

In this example, the two models with the lowest AIC have the poorest extrapolation, when compared to the true function. The model that performs best, but a long way from perfect, is the Weibull model, which had the highest AIC. The AIC or BIC criteria only uses the available data and in fact greater weight is given to earlier survival time data typically, as this is where the events are most dense. Thus a good fit within the range of the data may not lead to good extrapolation.

Plotting both the extrapolated survival and hazard functions together with the expected mortality rate is useful and recommended for any extrapolation. In this example, it clearly shows that the extrapolated functions are predicting a lower mortality rate in the later years than if the study population were disease free. From such plots it becomes obvious that none of the fitted models are suitable to assess overall mean survival and hence all should be rejected for this purpose.

**Figure 20: Plot of (a) survival and (b) hazard functions for various parametric and flexible parametric models up to the end of trial follow-up (3-years) for one realisation of Scenario 1**

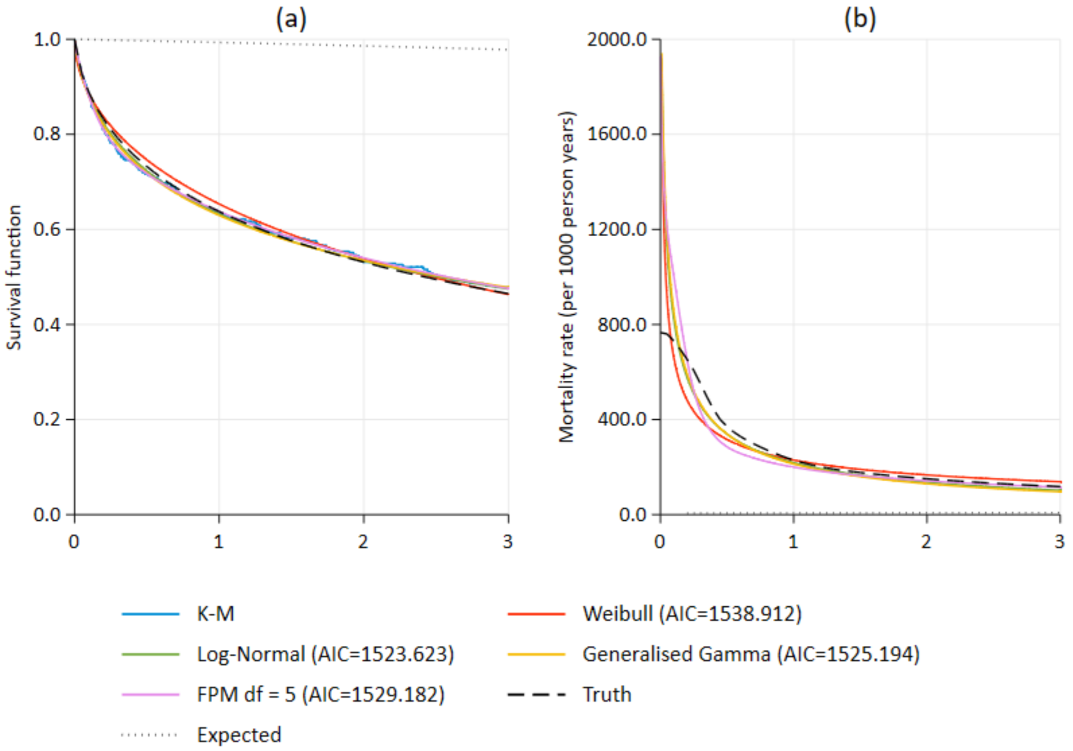
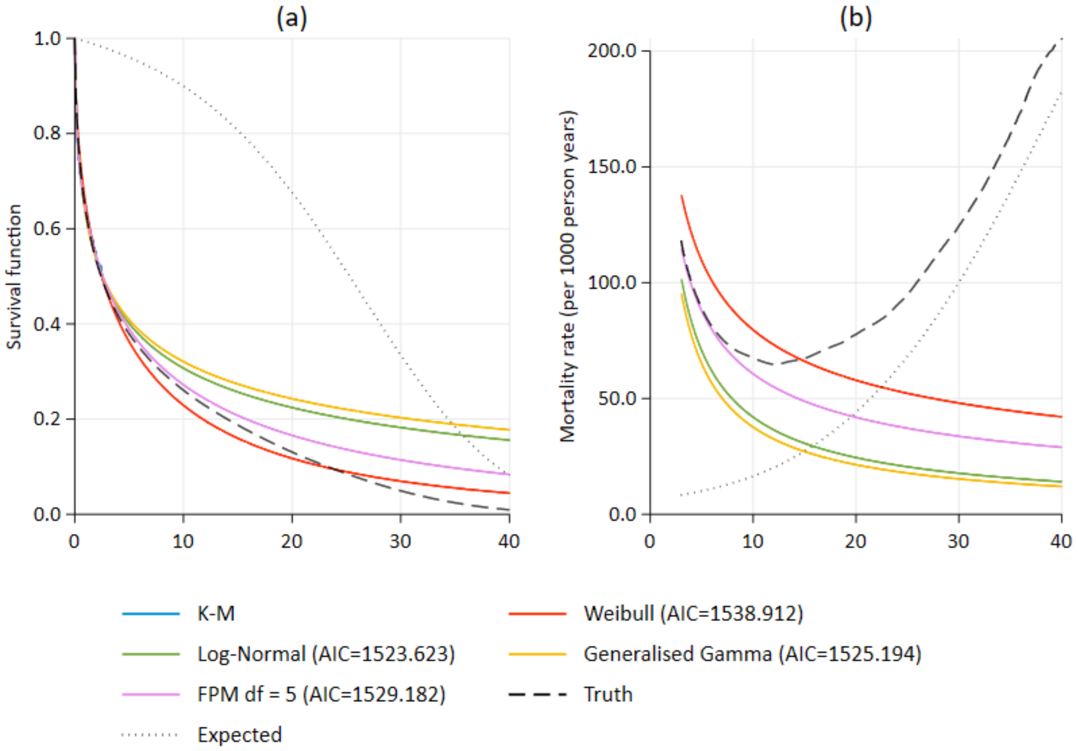


**Figure 21: Plot of (a) survival and (b) hazard functions for various parametric and flexible parametric models extrapolated up to 40-years for one realisation of Scenario 1**



## 6.4. INCORPORATING BACKGROUND MORTALITY INTO MODELLING MAY HELP WITH EXTRAPOLATION

This example uses the same dataset as the previous section (Section 5.2), but now compares FPMs with and without incorporating background mortality. The background mortality used are general English mortality rates for 2009. Since these are the rates that were used to generate mortality due to other causes in the simulation study, one would expect the simulation to perform well. As a sensitivity analysis, in Section 5.5 we show an example of using the “wrong” background mortality rates.

Figure 22 plots (a) the estimated marginal survival functions and (b) the marginal hazard functions up to the end of trial follow-up at 3-years. Within the range of the data there is good agreement between the two FPMs, the non-parametric Kaplan-Meier curve and the true survival function. There is also good agreement between the methods in terms of the hazard function; in fact the fitted functions cannot be distinguished on the plots.

Figure 23 plots (a) the estimated marginal survival functions and (b) the marginal hazard functions extrapolated to 40 years. The standard FPM model overestimates survival. The reason for this is clear through inspection of the hazard function, which continues to decrease. Comparing with the expected hazard in the general population shows that extrapolation of the standard FPM leads to a mortality rate that is lower than expected in the general population, so is clearly not sensible. Again, this shows the value of plotting the hazard together with a relevant expected mortality rate. In contrast to the standard FPM, the FPM that incorporates background mortality closely predicts the long-term survival and hazard functions.

**Figure 22: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality up to the end of trial follow-up (3-years) for one realisation of Scenario 1**

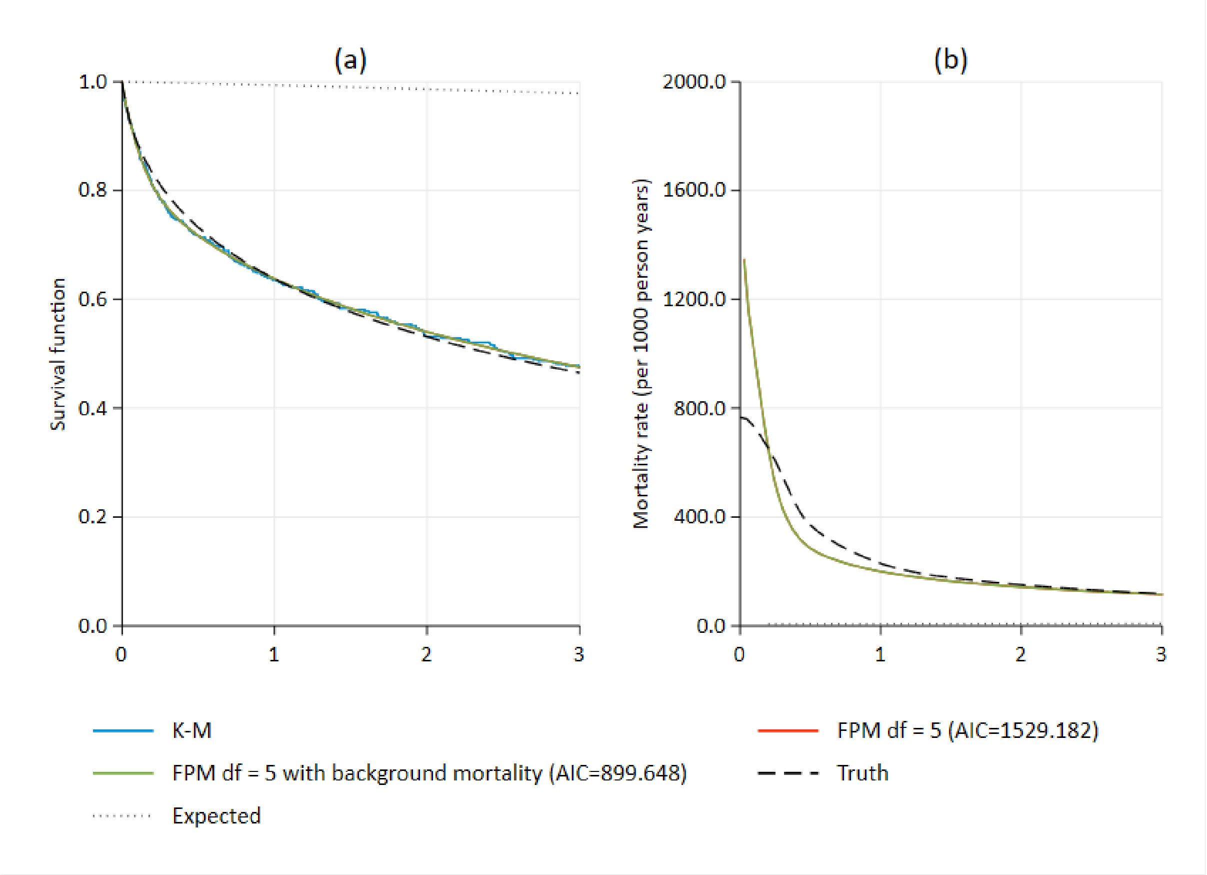
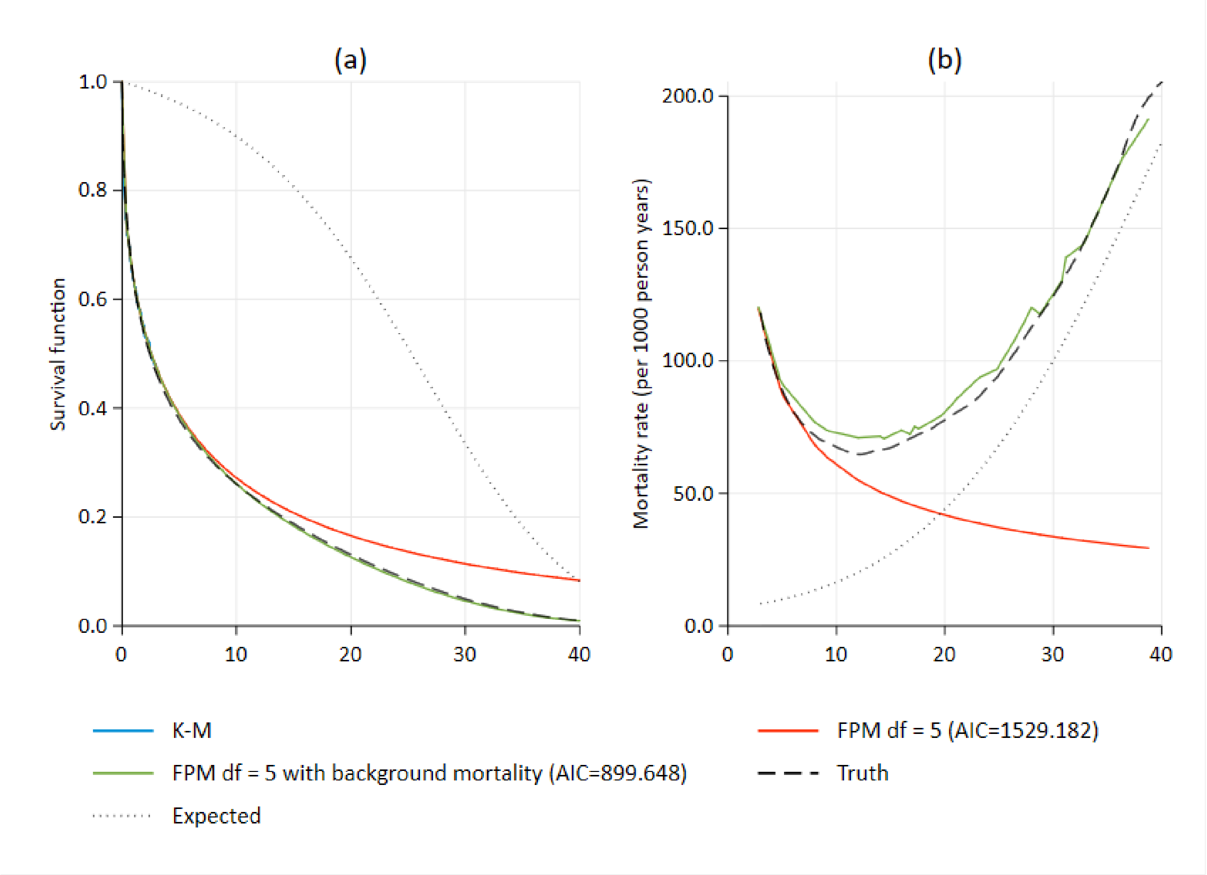


**Figure 23: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality extrapolated up to 40 years for one realisation of Scenario 1**



## 6.5. What might happen if incorrect background mortality rates are used?

Here we repeat the analysis from Section 5.4, but use incorrect background mortality rates. We use the rates for Females in Finland in 2000.

Figure 24 shows (a) the estimated marginal survival functions and (b) the marginal hazard functions up to the end of trial follow-up at 3-years. Within the range of follow-up using the incorrect expected survival rates has had little impact.

Figure 25 plots (a) the estimated marginal survival functions and (b) the marginal hazard functions extrapolated to 40 years. The expected mortality rates in Finland in 2000 were higher than in England in 2009 and this has led to the FPM that includes background mortality rates extrapolating less well than the version that included the correct background mortality rates. There is a slight underestimation of expected survival with this model, but it is still clearly better than the FPM that does not incorporating any background mortality rates.

**Figure 24: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality up to the end of trial follow-up (3-years) for one realisation of Scenario 1. Background mortality is misspecified**

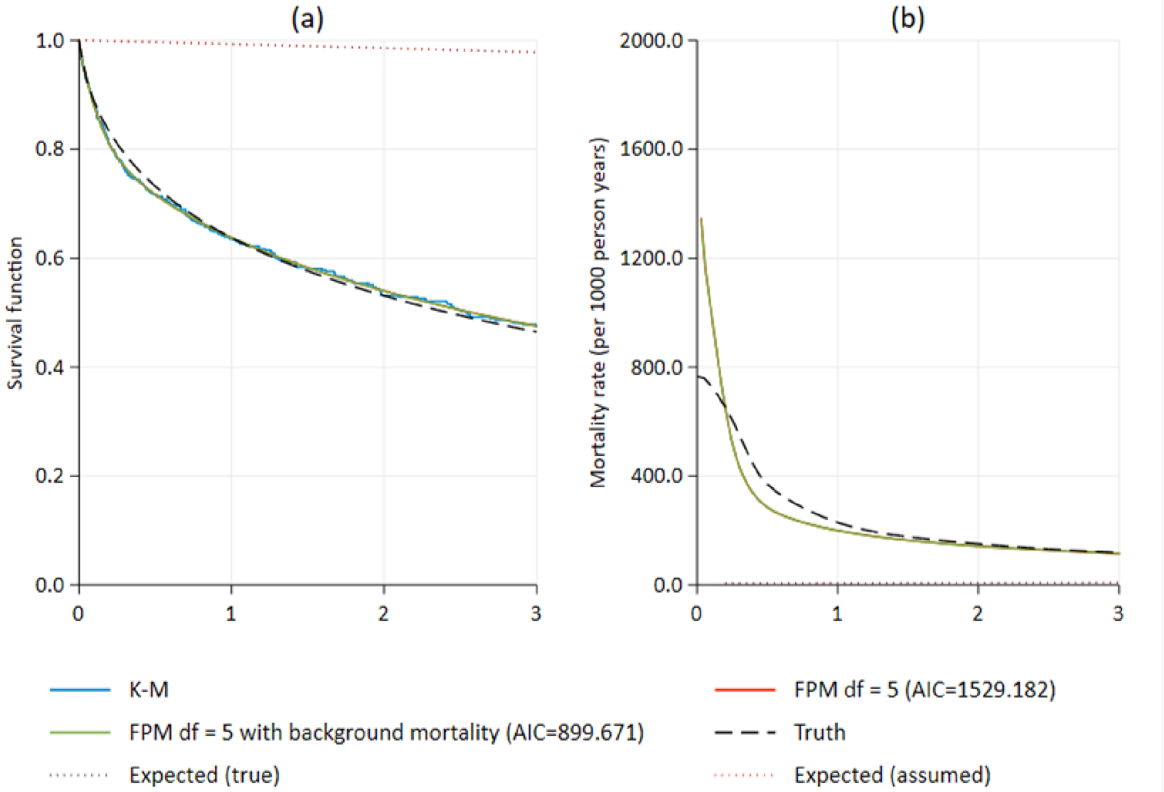
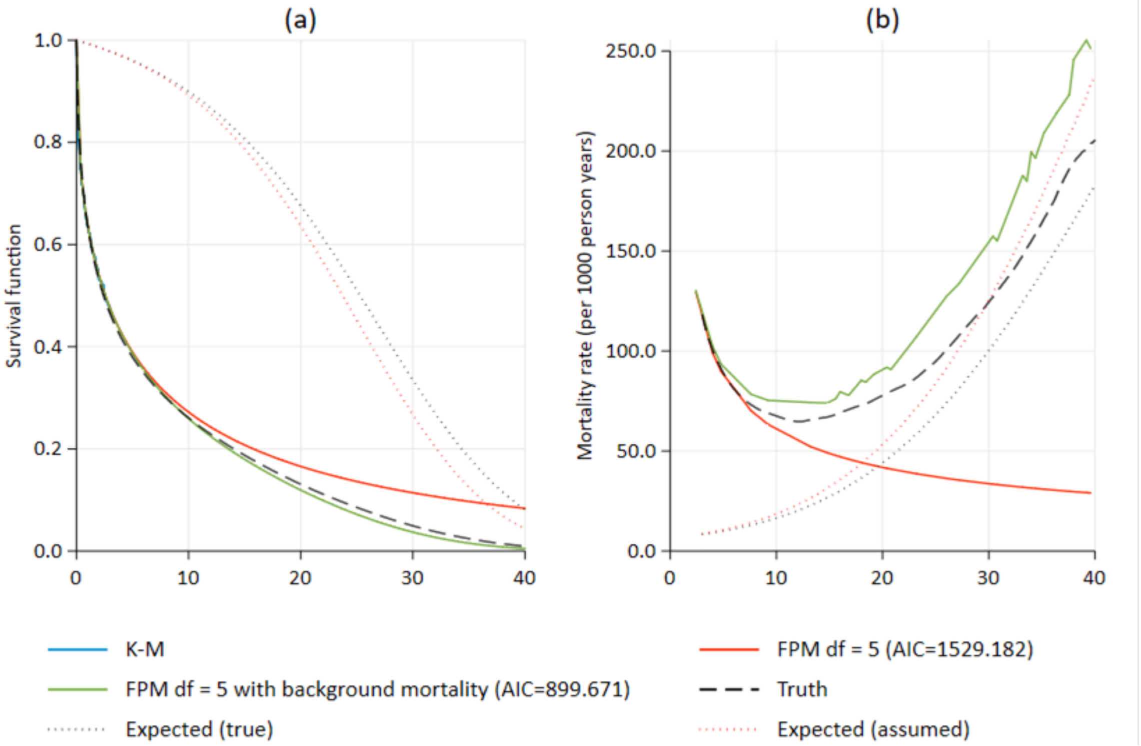


**Figure 25: Plot of (a) survival and (b) hazard functions for a flexible parametric model with and without incorporation of background mortality extrapolated up to 40 years for one realisation of Scenario 1. Background mortality is misspecified**



## 6.6. Having longer-term follow-up information is likely to be beneficial, but not always for all-cause approach

From the simulation results in Section 4, it is clear that many of the methods perform poorly for Scenario 3 (Figure 19). Given that the data are simulated, we can investigate applying the approaches in the case where follow-up was extended to 5 years, rather than 3 to see if the extrapolation is then improved.

Figure 26 shows the extrapolated survival functions using the simulated trial data with administrative censoring times of 3 and 5 years. True marginal survival is compared to two FPM predictions, one from an all-cause model, and the other from an excess mortality model with background mortality incorporated into the survival extrapolation. In the left-hand panel, both the FPM approaches with and without background mortality underestimate the true survival with 3 years of follow-up. Extending to 5 years in the right-hand panel improves the estimate for the FPM incorporating background mortality. This is because the true disease-specific hazard has a turning point after 3 years that can now be better captured in the excess mortality model. The FPM with no background mortality information to define the long-term hazard now overestimates the true survival in the long-term. Extending the follow-up to 5 years allows the cause-specific hazard to plateau appropriately meaning that the model using background mortality in the long-term now performs well – we are well capturing the shapes of both hazards. However, the extension to 5 years of follow-up has made the all-cause model worse in terms of the estimated mean survival time. Not incorporating the known long-term increase in hazards due to death from other causes can lead to an underestimation of the long-term hazard. Even with longer-term follow-up and lower levels of censoring it is important to ensure that hazard projections are credible.

The corresponding plots for the marginal hazard functions are shown in Figure 27. These plots allow a clearer understanding of the assumptions that are being made at varying points in follow-up, but must be interpreted in correspondence with the number/proportion of individuals still at risk. Models that are based on the fit within the range of the data alone are very unlikely to capture the true upturn in hazard due to population ageing, and the competing hazard of death due to other causes, which can be seen for the true hazard function.

**Figure 26: Plot of marginal survival for one realisation of Scenario 3, with varying follow-up lengths across the sub-panels. The grey shaded area shows the period where the survival functions is extrapolated (i.e. beyond the range of the trial follow-up).**

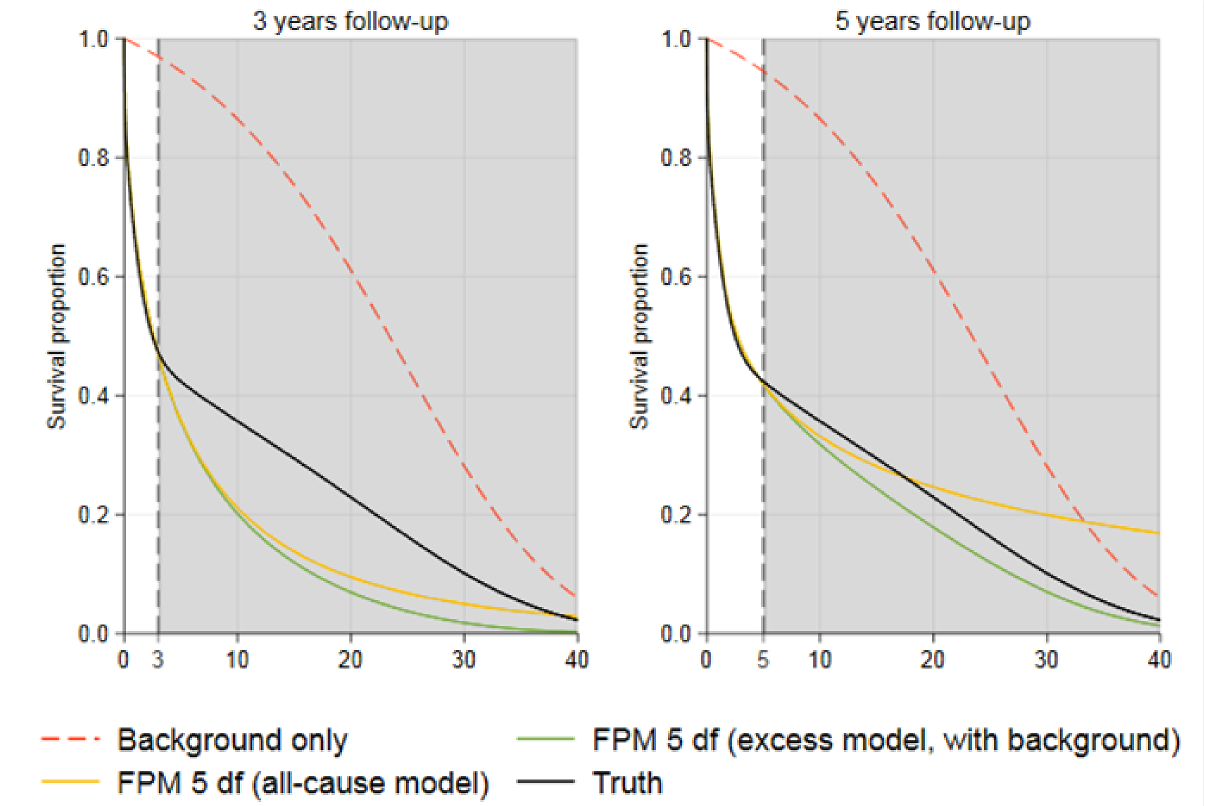
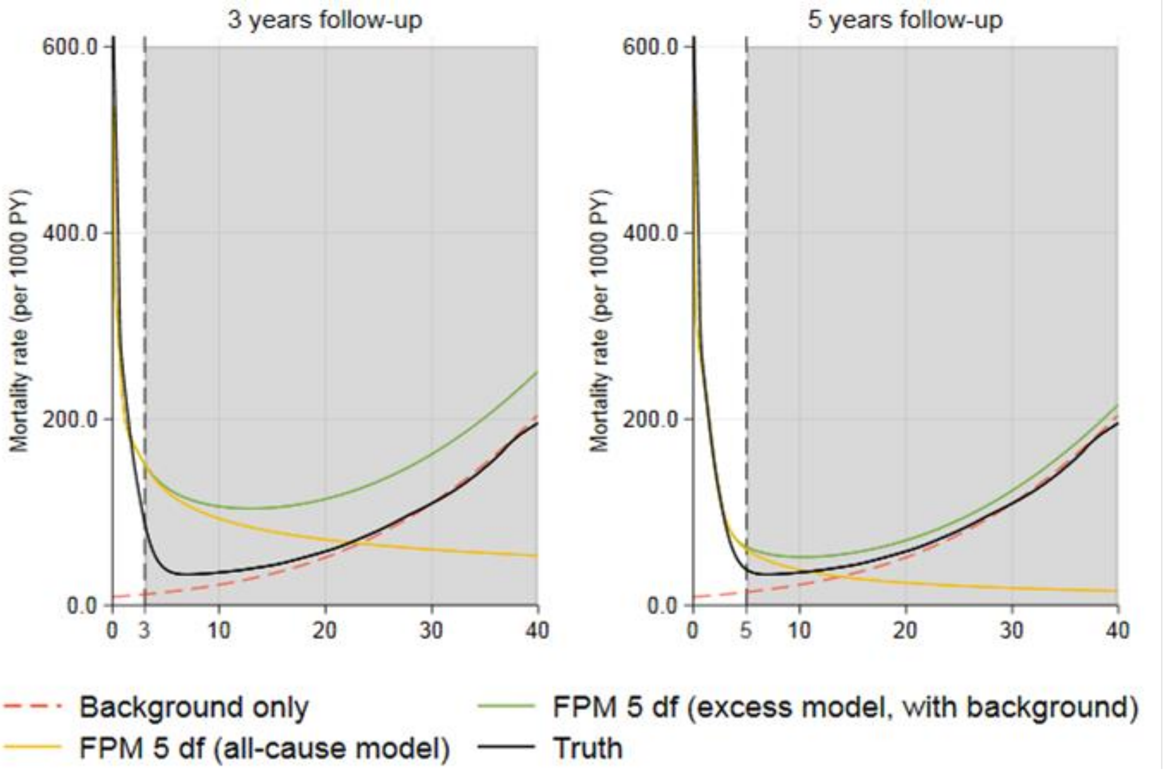


**Figure 27: Plot of marginal hazard function for one realisation of Scenario 3, with varying follow-up lengths across the sub-panels.**



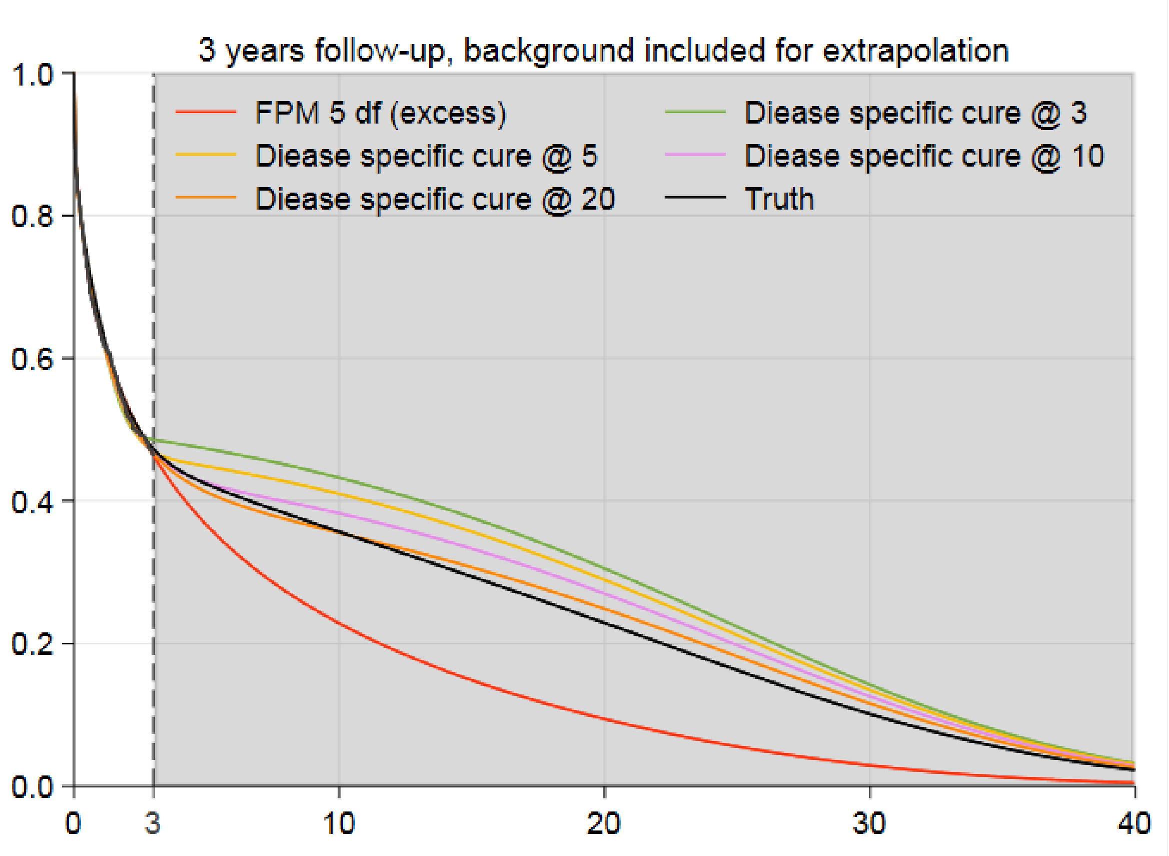
### 6.7. Altering the time to cure in order to use the population hazard only in the long-term

In the context of extrapolation, cure models are often used to allow the long-term mortality to be dominated by the population hazard. Flexible parametric cure models implement a time to cure directly, rather than having an asymptote at an infinite event time, and therefore, are very well suited to this context. In selecting a cure time exactly, one can allow the long-term hazard to be exactly that of the mortality of the general population or that of an appropriate disease register.

In Figure 28, true marginal survival is compared to various FPM predictions – one from an excess mortality model with background mortality incorporated into the survival extrapolation, and the others from FPM cure models with specified cure time-points. As can be seen, in this case, assuming that disease-specific mortality reaches 0 at a fixed point in follow-up time, performs better than allowing the standard FPM excess mortality model that incorporates background mortality. However, there is no fixed way to select the time to cure point, and in this particular simulation we know that the standard FPM which incorporates background mortality performs poorly because there is a turning point in the disease specific hazard after 3 years – this will usually not be known unless there is clinical knowledge to understand that this will likely be the case.

Despite this, this approach shows in principle how constraints on the disease-specific model, coupled with external data for other-cause mortality, can be used to capture complex hazard functions, but allow the longer-term hazard to be fully based on the external data hazard. Cure models in this context are useful for ensuring the long-term hazard is more fully defined by the external data used to arrive at the all-cause marginal survival.

**Figure 28: Plot of marginal hazard function for one realisation of Scenario 3, with varying assumed cure points**

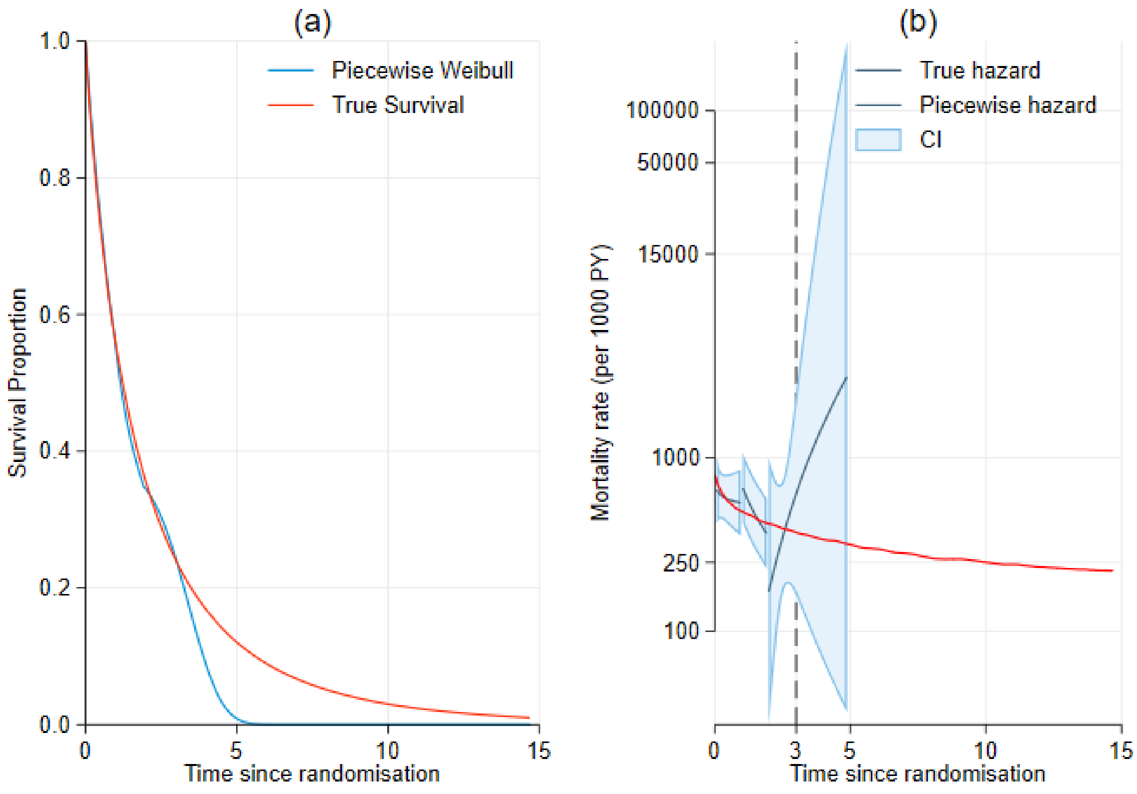


## 6.8. Piecewise models; Reduction of sample size at later cut-points

As mentioned in Sections 2 and 3, a piecewise modelling approach allows flexibility in terms of the hazard shapes that can be characterised, but there is likely to be an even more stark issue of sample size for the final piecewise section from which the extrapolation is made. Piecewise models could not be sensibly included in our simulation study, but here we demonstrate their application to one simulation from Scenario 1.

Figure 29 plots true survival and hazard functions and compares these to predictions from a piecewise model that fits 3 separate Weibull models to the 3 years of simulated data. The fit to the observed data is reasonable, but there is a great deal of uncertainty around the final piece of the model (demonstrated by the confidence intervals around the hazard function shown in panel (b) of Figure 29). The fitted model does not extrapolate well, and results in substantial bias. Care should be taken when using a piecewise approach - through studying the estimated hazard function, but also through illustrating the associated confidence intervals in tandem with being explicit about sample sizes.

**Figure 29: Plot of (a) survival and (b) hazard functions for one realisation of Scenario 1. Piecewise models**



# 7. DISCUSSION AND RECOMMENDATIONS

It is not the objective of this document to make recommendations on which types of survival model should or should not be used. Indeed, we have shown that there is no one type of survival model that will always produce the best survival extrapolations. In general, the more of the survival function left to predict (depending of course on prognosis and length of follow-up in the trial), the more scope there is to go wrong by applying a model with an inappropriate extrapolation. Because we can never know which model predicts most accurately, it is important to present models that incorporate a range of plausible assumptions about the long-term hazard, or to select models that appropriately allow for the differing mechanisms impacting on survival.

Often in oncology appraisals the trial data requiring extrapolation will be subject to treatment switching, i.e. patients in the control group will switch to the treatment under investigation at some point during the trial, e.g. on disease progression. Thus, adjustment for this must be made prior to undertaking extrapolation (see TSD16)[63]. In addition, an indirect comparison may also be required, for example using a population-adjusted approach (see TSD18)[12]. Thus, the impact of both of these factors on ultimate decisions and cost-effectiveness estimates may also have to be explored in combination with sensitivity to the methods described in this document.

In this document we have attempted to describe the characteristics of different types of complex survival model, and to demonstrate their potential performance in a range of plausible and realistic scenarios. In addition, we have sought to demonstrate the kinds of analyses and plots that may be helpful in showing which models might be appropriate, and what the alternative models actually project. To this end, we make the following general and model-specific recommendations together with recommendations for further research.

## 7.1. GENERAL RECOMMENDATIONS:

I. **Plotting predicted survival and hazard functions.** Fitted and extrapolated hazard and survival functions should always be presented. Exploring model assumptions and implications on the hazard scale is particularly important because this is the scale which models are estimated on, and often the scale upon which treatment effects are assumed to act. This also clearly demonstrates what is being assumed about the hazard function over time, which is an easier scale on which to visualise and conceptualise risk changing through time as opposed to a change in gradient for a cumulative measure. A justification should also be given to explain why the projected hazard and survival functions are credible.

II. **Plotting expected (general population) survival and hazard functions.** This can aid understanding of whether the assumed hazard and survival functions are credible.

III. **Incorporation of background mortality.** Incorporating background mortality into survival models is recommended because it helps avoid extremely implausible projections. This is true for standard parametric models, FPMs, mixture models, landmark models and piecewise models and is *essential* for cure models. Background mortality rates should either be incorporated when making the extrapolation, or used as a sense-check when plotting the marginal survival, and particularly the marginal hazard functions that have been extrapolated. National mortality rates stratified by age, sex and calendar year may be used. However, if study subjects have more comorbidities than the general population, expected rates will be underestimated. Attempts could be made to account for this, but even if the background mortality rates used are inappropriate, the extrapolation is likely to be better than that associated with a model that does not take into account background mortality. If the excess mortality rates are assumed to be zero (when predicted hazard rates are the same as background mortality rates) then the time point this occurs should be stated.

IV. **Incorporation of other external information.** Other external information, such as registry data, may be useful to incorporate within survival models. However, research is ongoing in this area and we cannot make firm recommendations. If relevant registry data are identified, relative comparisons between the trial population, the disease population of interest, and the registry population should be made and, if possible, registry data from the most relevant patients should be used. Consideration of available external data sources should be done at an early stage, and analyses incorporating this information should be pre-specified in analysis plans.

V. **Treatment effects.** In this document we have concentrated on extrapolation for a single group of patients. When extrapolating for two groups of patients, hazard functions for both groups should be plotted together with the implied treatment effect (whether a proportional treatment effect is assumed, or whether survival models are fitted separately to treatment groups). External information (e.g. registry data) is likely to be of most use for control group extrapolations. Extrapolation for the experimental group should follow the same principles described above (including background mortality information) but should also incorporate sensitivity analyses around long-term treatment effects.

### 7.2. MODEL-SPECIFIC RECOMMENDATIONS:

VI. **Standard parametric models.** Standard parametric models can provide reasonable extrapolations if long-term hazards are expected to follow simple shapes. However this is frequently likely not to be the case. When standard parametric models are used background mortality rates and/or other relevant external information should always be considered for incorporation.

VII. **Flexible parametric models.** FPMs are very likely to fit the observed survival data well, but may not extrapolate appropriately. When FPMs are used, background mortality rates and/or other relevant external information should always be considered for incorporation.

VIII. **Mixture models.** Mixture models may be intuitively appealing, but frequently suffer from lack of convergence and can be mis-interpreted. Mixture models should be used with extreme caution – FPMs are likely to represent a more reliable option for modelling complex hazard functions.

IX. **Landmark models.** If landmark models are used care should be taken to justify group categorisations, and consideration should be given to the sensitivity to the landmark time-point chosen. Uncertainty around models fitted to different categories should be clearly demonstrated through the use of plots of modelled hazards and survival, with confidence intervals shown. Background mortality rates and/or other relevant external information should be included within whichever models are fitted to the different categories.

X. **Piecewise models.** If piecewise models are used care should be taken to justify the time-points used. As for all models, observed and predicted hazard and survival plots should be presented. Sensitivity to time-points chosen should be explored. Uncertainty around the models fitted to the different time-points should be clearly demonstrated through the use of plots of modelled hazards and survival, with confidence intervals shown. Background mortality rates and/or other relevant external information should be included within each of the models fitted, but particularly importantly for the model fitted to the final segment of survival.

XI. **Cure models.** Cure models may be useful when an assumption of cure is reasonable – however, it is the incorporation of background mortality that is important rather than the fact that the model is a “cure model”. Hence, cure models hold few advantages compared to FPMs that are fitted incorporating background mortality using a relative survival or excess mortality framework. Cure models perform poorly if they are applied and an assumption of “cure” is not reasonable. Hence, it is likely to be more appropriate to use FPMs incorporating background mortality instead of cure models if the existence of a “cure” is uncertain. Estimating the cure fraction is prone to high levels of uncertainty, particularly when sample sizes are small. If a cure model is used, evidence should be provided to justify the credibility of the cure fraction estimated (e.g. by comparing to response rates, as well as by exploring earlier phase trials with longer follow-up, and using biological information on the mechanism of action of the treatment and the nature of the disease). When using a cure model, the hazard and survival functions predicted (and expected) for cured and uncured populations should both be presented.

## 7.3. Recommendations for Research:

I. We have presented a simulation study investigating the performance of different survival models in different realistic circumstances. However, simulation studies cannot be exhaustive. It would be useful to explore other scenarios.

II. There are many ways in which external data sources could be incorporated into survival analyses and extrapolation. Further research is required exploring the most appropriate for specific situations. Bayesian methods would appear to offer a means by which both expert opinion and external data sources, together with model uncertainty, could be explored and integrated into health technology assessment.

III. The approaches that we have outlined have largely tried to capture and extrapolate the marginal hazard and survival functions without trying to compartmentalise the mechanisms driving changes in these functions. Some of the approaches – such as utlising external data start to elucidate the competing mechanisms that drive the overall marginal functions. Methods that directly model the competing components and are tailored to the mechanisms and covariate effects governing the all-cause survival function could be explored in future research, and would allow sensitivity analysis to the assumptions made.

IV. A strong determinant of the lifetime benefits of a treatment will be the assumed long-term treatment effect. Approaches that directly attempt to model the long-term relative treatment effect, and also assess deviations from the assumptions surrounding this would offer an alternative approach to extrapolation. Extrapolating relative treatment effects could involve borrowing information from similar drug classes and other longer-term clinical trial follow-up, and/or eliciting expert opinion. Further research, and evaluation, is required in order to explore the viability of this approach.

V. Trial populations may not fully reflect the target population of interest for which we wish to make our decision. We have taken a largely trial-based perspective for estimation of the mean survival, but methods to reweight to a target population appropriately when performing extrapolations deserve further attention, research and evaluation.

# APPENDIX

## A.1 STATA CODE TO GENERATE SIMULATION STUDY DATA

See the Simulation folder in the TSD_Stata_Files download available at: [*to be filled in once on DSU website*]

## A.2 STATA CODE USED TO UNDERTAKE ANALYSES

See the Example folder in the TSD_Stata_Files download available at: [*to be filled in once on DSU website*]

## A.3 STATA CODE TO CALCULATE MARGINAL EXPECTED SURVIVAL AND HAZARD

See the stexpect3 folder in the TSD_Stata_Files download, and the corresponding worked through code in the Example folder.

## A.4 SIMULATION STUDY RESULTS

See the separate Simulation_results.pdf file for the tables and figures relating to the full simulation results.

# REFERENCES

1. Latimer, N., *NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data*, in *Report by the Decision Support Unit*, A. Wailoo, Editor. 2011: Available from http://www.nicedsu.org.uk
2. Latimer, N.R., *Survival Analysis for Economic Evaluations Alongside Clinical Trials—Extrapolation with Patient-Level Data:Inconsistencies, Limitations, and a Practical Guide.* Medical Decision Making, 2013. **33**(6): p. 743-754.
3. Bullement, A., N.R. Latimer, and H. Bell Gorrod, *Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study.* Value Health, 2019. **22**(3): p. 276-283.
4. Ouwens, M.J.N.M., et al., *Estimating Lifetime Benefits Associated with Immuno-Oncology Therapies: Challenges and Approaches for Overall Survival Extrapolations.* PharmacoEconomics, 2019. **37**(9): p. 1129-1138.
5. Othus, M., et al., *Accounting for Cured Patients in Cost-Effectiveness Analysis.* Value Health, 2017. **20**(4): p. 705-709.
6. Tan, S.H., K.R. Abrams, and S. Bujkiewicz, *Bayesian Multiparameter Evidence Synthesis to Inform Decision Making: A Case Study in Metastatic Hormone-Refractory Prostate Cancer.* Medical decision making : an international journal of the Society for Medical Decision Making, 2018. **38**(7): p. 834-848.
7. National Institute for Health and Care Excellence (NICE). *Lenvatinib with everolimus for previously treated advanced renal cell carcinoma*. NICE Guidance TA498 2018 Available at https://www.nice.org.uk/guidance/ta498/documents/html-content-2].
8. National Institute for Health and Care Excellence (NICE). *Avelumab for treating metastatic Merkel cell carcinoma* NICE Guidance TA517 2018 Available at: https://www.nice.org.uk/guidance/ta517/documents/html-content-2].
9. National Institute for Health and Care Excellence (NICE). *Brentuximab vedotin for treating relapsed or refractory systemic anaplastic large cell lymphoma*. NICE Guidance TA478 2017 Available at: https://www.nice.org.uk/guidance/ta478/documents/html-content-2].
10. National Institute for Health and Care Excellence (NICE). *Cabozantinib for previously treated advanced renal cell carcinoma* NICE Guideance TA463 2017 Available at: https://www.nice.org.uk/guidance/ta463/documents/html-content-3].
11. National Institute for Health and Care Excellence (NICE). *Nivolumab for previously treated locally advanced or metastatic squamous nonsmall-cell lung cancer*. NICE Guidance TA483 2016 Available at: https://www.nice.org.uk/guidance/ta483/documents/appraisal-consultation-document-2].
12. Phillippo, D.M., et al., *NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE*, in *Report by the Decision Support Unit*, A. Wailoo, Editor. 2016: Available from http://www.nicedsu.org.uk
13. Abrams, K., et al., *Propensity Weighting and Extrapolation in Non Small Cell Lung Cancer. Work Package 1, IMI GetReal*

2016: https://www.imi-getreal.eu/Portals/1/Documents/01%20deliverables/Deliverable%201.5%20and%201.6%20Combined%20Report%20-%20NSCLC_webversion.pdf.

14. Bell Gorrod, H., et al., *A Review of Survival Analysis Methods Used in NICE Technology Appraisals of Cancer Treatments: Consistency, Limitations, and Areas for Improvement.* Medical Decision Making, 2019. **39**(8): p. 899-909.
15. Royston, P. and M.K. Parmar, *Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects.* Stat Med, 2002. **21**(15): p. 2175-97.
16. Rutherford, M.J., M.J. Crowther, and P.C. Lambert, *The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study.* Journal of Statistical Computation and Simulation, 2015. **85**(4): p. 777-793.
17. Bower, H., et al., *Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study.* Communications in Statistics - Simulation and Computation, 2019: p. 1-17.
18. Syriopoulou, E., Mozumder, S.I., Rutherford, M.J., Lambert, P.C., *Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models.* Cancer Epidemiology, 2019. **58**: p. 17-24.
19. Royston, P., Lambert, P.C., *Flexible parametric survival analysis in Stata: Beyond the Cox model* Stata Press, 2011.
20. Andersson, T.M.-L., et al., *Estimating the loss in expectation of life due to cancer using flexible parametric survival models.* Statistics in Medicine, 2013. **32**(30): p. 5286-5300.
21. Liu, X.R., Pawitan, Y., Clements, M., *Parametric and penalized generalized survival models*

Statistical Methods in Medical Research, 2018. **27(5)**: p. 1531-1546.

22. Demiris, N., D. Lunn, and L.D. Sharples, *Survival extrapolation using the poly-Weibull model.* Statistical Methods in Medical Research, 2011. **24**(2): p. 287-301.
23. Demiris, N., D. Lunn, and L.D. Sharples, *Survival extrapolation using the poly-Weibull model.* Statistical methods in medical research, 2015. **24**(2): p. 287-301.
24. Farewell, V.T., *Mixture Models in Survival Analysis: Are They Worth the Risk?* The Canadian Journal of Statistics / La Revue Canadienne de Statistique, 1986. **14**(3): p. 257-262.
25. McLachlan, G.J., S.X. Lee, and S.I. Rathnayake, *Finite Mixture Models.* Annual Review of Statistics and Its Application, 2019. **6**(1): p. 355-378.
26. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooneym M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., Verweij, J., *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).* European Journal of Cancer, 2009. **45(2)**: p. 228-247.
27. Anderson, J.R., K.C. Cain, and R.D. Gelber, *Analysis of survival by tumor response.* Journal of Clinical Oncology, 1983. **1**(11): p. 710-719.
28. Dafni, U., *Landmark analysis at the 25-year landmark point.* Circ Cardiovasc Qual Outcomes, 2011. **4**(3): p. 363-71.
29. National Institute for Health and Care Excellence (NICE). *Everolimus with exemestane for treating advanced breast cancer after endocrine therapy*. NICE

Guidance TA421 2016 Available at: https://www.nice.org.uk/guidance/ta421/documents/html-content-3].

30. Casellas, J., *Bayesian inference in a piecewise Weibull proportional hazards model with unknown change points.* Journal of Animal Breeding and Genetics, 2007. **124**(4): p. 176-184.
31. Coelho-Barros, E.A., et al., *Bayesian Inference For The Segmented Weibull Distribution.* 2019, 2019. **42**(2): p. 19.
32. Friedman, M., *Piecewise Exponential Models for Survival Data with Covariates.* Ann. Statist., 1982. **10**(1): p. 101-113.
33. Gelber, R.D., A. Goldhirsch, and B.F. Cole, *Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments.* Controlled Clinical Trials, 1993. **14**(6): p. 485-499.
34. National Institute for Health and Care Excellence (NICE). *Nivolumab for treating squamous cell carcinoma of the head and neck after platinum-based chemotherapy*. NICE Guidance TA490 2017 Available at: https://www.nice.org.uk/guidance/ta490/documents/html-content-2].
35. National Institute for Health and Care Excellence (NICE). *Abiraterone for treating metastatic hormone-relapsed prostate cancer before chemotherapy is indicated*. NICE Guidance TA387 2016 Available at: https://www.nice.org.uk/guidance/ta387/documents/html-content-2].
36. Gong, Q. and L. Fang, *Asymptotic properties of mean survival estimate based on the Kaplan–Meier curve with an extrapolated tail.* Pharmaceutical Statistics, 2012. **11**(2): p. 135-140.
37. Bagust, A., Beale, S., *Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach.* Medical Decision Making, 2014. **34(3)**: p. 343-51.
38. National Institute for Health and Care Excellence (NICE). *Cabazitaxel for hormone-relapsed metastatic prostate cancer treated with docetaxel*. NICE Guideline TA391 2016 Available at: https://www.nice.org.uk/guidance/ta391/documents/html-content-2].
39. National Institute for Health and Care Excellence (NICE). *Pembrolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum-containing chemotherapy*. NICE Guidance TA519 2018 Available at: https://www.nice.org.uk/guidance/ta519/documents/html-content-2].
40. Davies, A., Briggs, A., Schneider, J. , *The ends justify the mean: outcome measures for estimating the value of new cancer therapies.* Health Outcomes Research in Medicine, 2012. **3**: p. e25–236.
41. National Institute for Health and Care Excellence (NICE). *Ipilimumab for previously treated advanced (unresectable or metastatic) melanoma*. NICE Guidance TA268 2012 Available at: https://www.nice.org.uk/guidance/ta268/documents/melanoma-stage-iii-or-iv-ipilimumab-final-appraisal-determination].
42. Boag, J.W., *Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy.* Journal of the Royal Statistical Society: Series B (Methodological), 1949. **11**(1): p. 15-44.
43. Sposto, R., *Cure model analysis in cancer: an application to data from the Children's Cancer Group.* Stat Med, 2002. **21**(2): p. 293-312.
44. Andersson, T.M.L., et al., *Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models.* BMC Medical Research Methodology, 2011. **11**(1): p. 96.

45. Yu, X.Q., De Angelis, R., Andersson, T.M., Lambert, P.C., O'Connell, D.L., Dickman, P.W., *Estimating the proportion cured of cancer: Some practical advice for users.* Cancer Epidemiology, 2013. **37(6)**: p. 836-842.
46. Dickman, P.W., and Adami, H-O., *Interpreting trends in cancer patient survival.* Journal of International Medicine, 2006. **260(2)**: p. 103-117.
47. Morris, D.E., J.E. Oakley, and J.A. Crowe, *A web-based tool for eliciting probability distributions from experts.* Environmental Modelling & Software, 2014. **52**: p. 1-4.
48. Spiegelhalter, D.J., K.R. Abrams, and J.P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Vol. DOI:10.1002/0470092602 2004: Wiley.
49. Baio, G., *survHE: Survival analysis for health economic evaluation and cost-effectiveness modelling. .* Journal of Statistical Software, 2020(In Press).
50. Soares, M.O., et al., *Experiences of Structured Elicitation for Model-Based Cost-Effectiveness Analyses.* Value in Health, 2018. **21**(6): p. 715-723.
51. Brard, C., et al., *Bayesian survival analysis in clinical trials: What methods are used in practice?* Clinical Trials, 2016. **14**(1): p. 78-87.
52. Grigore, B., et al., *Methods to Elicit Probability Distributions from Experts: A Systematic Review of Reported Practice in Health Technology Assessment.* PharmacoEconomics, 2013. **31**(11): p. 991-1003.
53. Guyot, P., et al., *Extrapolation of Survival Curves from Cancer Trials Using External Information.* Med Decis Making, 2017. **37**(4): p. 353-366.
54. Soikkeli, F., et al., *Extrapolating Survival Data Using Historical Trial-Based a Priori Distributions.* Value Health, 2019. **22**(9): p. 1012-1017.
55. Ibrahim, J., Chen, M., Sinha, D.,, *Bayesian Survival Analysis*. 2001, https://doi.org/10.1007/978-1-4757-3447-8: Springer, New York, NY.
56. Singpurwalla, N.D., Song, M.S.,, *The Analysis of Weibull Lifetime Data Incorporating Expert Opinion. In:*, in *Probability and Bayesian Statistics.* , V. R., Editor. 1987, Springer, Boston, MA
57. Cope, S., et al., *Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia.* BMC Medical Research Methodology, 2019. **19**(1): p. 182.
58. Walsh, D.P., et al., *Using expert knowledge to incorporate uncertainty in cause-of-death assignments for modeling of cause-specific mortality.* Ecology and Evolution, 2018. **8**(1): p. 509-520.
59. Jackson, C., Stevens J., Ren, S., Latime,r N., Bojke, L., Manca, A., Sharples, L., *Extrapolating survival from randomized trials using external data: A review of methods.* Medical Decision Making 2017. **37(4)**: p. 377-390. .
60. Morris, T.P., I.R. White, and M.J. Crowther, *Using simulation studies to evaluate statistical methods.* Stat Med, 2019. **38**(11): p. 2074-2102.
61. Hougaard, P., *Frailty models for survival data.* Lifetime Data Analysis, 1995. **1**(3): p. 255-273.
62. Crowther, M.J. and P.C. Lambert, *Simulating complex survival data.* Stata Journal, 2012. **12**(4): p. 674-687.
63. Latimer, N. and K. Abrams, *NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching*, in *Report by the Decision Support Unit*, A. Wailoo, Editor. 2014: Available from http://www.nicedsu.org.uk