Time series methods for extrapolating survival data in health technology assessment

Benjamin Christopher Kearns

A thesis submitted for the degree of Doctor of Philosophy.

Supervised by Professor Matt Stevenson, Dr Kostas Triantafyllopoulos, and Professor Andrea Manca.

Health Economics and Decision Science, School of Health and Related Research

Submitted October 2020

Abstract

Extrapolating survival data is an important task in health technology assessment (HTA). Current approaches can lack flexibility and there is a tension between using all the data and restricting analyses to the more recent observations. Dynamic survival models (DSMs) exploit the temporal structure of survival data, are very flexible, have interpretable extrapolations, and use all the data whilst providing more weight to more recent observations when generating extrapolations. DSMs have not previously been used in HTA; this thesis evaluated the performance and usefulness of dynamic models in this context.

Extensive simulation studies compared DSMs with both current practice and other flexible (emerging practice) models. Results indicated that, compared with current practice, DSMs can more accurately model the data, providing improved extrapolations. However, with small sample sizes or short follow-up there was a danger of providing worse extrapolations. Of emerging practice models, spline-based models often had similar performance to DSMs whilst fractional polynomials provided very poor extrapolations. Two novel extensions to DSMs were developed to incorporate external data in the form of relative survival and cure models. Both extensions helped to reduce the variation in extrapolations. Dynamic cure models were assessed in a simulation study and provided good extrapolations that were robust to model misspecification. A case-study demonstrated that extrapolations from an interim analysis can be poor for all the methods considered when the observed data is not representative of the future. A case-study demonstrated the feasibility of using DSMs in HTA, and an extension to incorporate time-varying treatment effects.

DSMs should be considered as a potential method when analysing and extrapolating survival data. These flexible models and their extensions show promise but have the danger of providing poor extrapolations in data-poor scenarios. More research is needed into identifying situations when use of these should be the default approach in HTA.

Acknowledgements and funding

Firstly a huge thank you to my three supervisors: Matt Stevenson, Kostas Triantafyllopoulos, and Andrea Manca. Your guidance, ideas and support have been immensley useful during this fellowship. Thank you also to my wife Rebecca, daughter Emilia and dog Elsie for your happiness, company and putting up with a few too many late nights. Thank you to Nick Latimer and Monica Hernandez Alava for doing my confirmation review, and the useful feedback and conversations during the course of the fellowship. Also thanks to Mark Clowes and Lesley Uttley for helping with the literature reviews. There are a huge number of other people in ScHARR who have helped along the way with this work. Particular thanks go to Kate Ren, Alan Brennan, John Stevens, Tracey Young, Amanda Lane, and Samaira Khan. I would also like to thank the members of my public advisory group for their time and enthusiasm: Chris Unwin, Irina Sidorova, John Kay, Tracy Green, Debjani Chatterjee, Jacqui Gath, and Ellen Paling. Your input into this project has been greatly appreciated.

Ben Kearns' time was funded by the National Institute for Health Research (NIHR) Doctoral Research Fellowship (DRF-2016-09-119) 'Good practice guidance for the prediction of future outcomes in health technology assessment'. The Health Economics and Outcome Measurement Theme of the NIHR Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (www.clahrc-yh.nihr.ac.uk) supported Ben Kearns in the preparation of this work. The views expressed are those of the authors, and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health and Social Care.

Dissemination

At the time of submission the following publications have arisen from this fellowship (hopefully with more to come):

1. Kearns B, Stevenson M, Triantafyllopoulos K, Manca A (2019) *Generalized linear models for flexible parametric modeling of the hazard function.* Medical Decision Making 39(7):867-878
2. Kearns B, Stevens J, Ren S, Brennan A (2019) *How uncertain is the survival extrapolation? A study of the impact of different parametric survival models on extrapolated uncertainty about hazard functions, lifetime mean survival and cost effectiveness.* PharmacoEconomics 38: 193–204
3. Bell Gorrod H, Kearns B, Stevens J, *et al.* (2019) *A review of survival analysis methods used in NICE technology appraisals of cancer treatments: Consistency, limitations and areas for improvement.* Medical Decision Making 39 (8): 899-909
4. Kearns B. (*In press*). *Hazard function modelling* in Wiley StatsRef-Statistics Reference Online. editors: Longford, Davidian, Kenett, Molenberghs.

I have presented parts of this work at the following national and international conferences:

- ISPOR Europe 2019: two poster presentations.
- OR61 Annual Conference 2019: two oral presentations.
- Young Statistican's Meeting 2019: oral presentation.

- ScHARR PGR Conference 2019: oral and poster presentations.

- Research Students' Conference 2018: oral and poster presentations.

To support this fellowship I have supervised a number of MSc dissertations on similar themes. The results from one of these will be a podium presentation at ISPOR Europe 2020: *Structural uncertainty in survival extrapolation: exploring the impact of four model averaging methods and adjusting for data maturity* by Chloe Hardern. I have also presented parts of this work to a number organisations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The focus of this thesis is on using time series methods to predict future survival outcomes within health technology assessment (HTA). These predictions of the future are commonly referred to as *extrapolations* within the HTA literature, or *forecasts* within the statistical literature. Predicting the future is difficult as by definition there is no available evidence at the time of the prediction to assess the accuracy of these estimates. Predicting the future is also often a key task within HTA as these extrapolations can influence whether or not potentially life-saving treatments are made available to patients. Hence it is important that the best possible statistical methods are used to generate extrapolations. Time series methods are routinely used for forecasting in many areas but they have not currently been used for the extrapolation of survival data in HTA. The main novel contribution of this thesis is to develop and assess the use of time series methods for predicting future survival to inform HTAs. This contribution includes the identification and development of suitable methods, along with a comprehensive assessment of their performance.

This chapter provides key introductory material for the thesis. It begins with a background to HTA and why extrapolations of survival outcomes are an important part of this process. The motivation for considering time series methods is then described. This chapter concludes by outlining the aims of this thesis, and how these are achieved by the subsequent chapters. In general, the writing style uses a passive voice. The exception is where I wish to emphasise my value added (novel) contributions to this thesis. In particular, this thesis forms part of a wider doctoral fellowship. As such, some of the work that I have undertaken on extrapolation for the fellowship is beyond the scope of this thesis. This includes two published first-author papers. One paper was designed to highlight the methods used in this thesis to a general medical audience [1]. The analyses performed for this publication have since been superseded by more complex work which is discussed in this thesis. This publication includes important methodological details, and so the whole publication is reproduced as Appendix 3. The second paper examined the impact on uncertainty during the extrapolated phase when using survival models [2]. A full treatment of uncertainty in both survival extrapolation and HTA is beyond the scope of this thesis but it is briefly explored in Section 6.4.2. These papers, and other examples, are discussed at appropriate parts of this thesis.

1.1 Health technology assessment

A health technology is any technology that can be used to improve population health. Ideally, any health technology that has been proven to work (i.e. is clinically effective) would be made available via a healthcare system such as the National Health Service (NHS) in England. However, healthcare resources are limited. This includes limited funds, staff and facilities. Because of this scarcity of resources, not all clinically effective health technologies can be funded. Instead, a framework is required to decide which health technologies should be funded or prioritised, given the resource constraints. HTA is concerned with the scientific evaluation of health technologies. It has two main components: *clinical effectiveness*, which evaluates the impact of a technology on health compared to existing alternatives; and *cost-effectiveness*, which considers the trade-off between the impact on individuals' health and the costs to the healthcare system: does the technology represent value for money [3]? HTA follows the principles of evidence-based medicine; decisions are informed by evidence that is identified via a systematic, transparent and explicit process [4].

Two key outcomes in HTA are the costs to the healthcare system and benefits to patients. Costs include both the cost of delivering the health technology and any long-term costs such as those for disease management, re-interventions, or treating side effects. A common measure of patient benefit is *quality adjusted life years* (QALYs) [5]. These weight remaining life expectancy (survival) by the quality of that life. These weights are known as utilities and are typically scaled such that a utility of one is equivalent to being in perfect health, whilst a utility of zero is equivalent to being dead. There are two primary advantages of using QALY. The first is that they incorporate both quantity and quality of life into a single summary statistic. The second is that they provide a common outcome by which different technologies may be compared. As such, the principles of HTA may be applied to any health technology. Examples of HTA include drugs for cancer [6], screening programmes [7], service (pathway) redesign [8] and alternative clinical practices [9].

If evidence were available on the costs and QALYs for every intervention, then the choice of which interventions to fund would reduce to a constrained optimization problem, where the constraint is the available healthcare budget (costs) and optimisation is performed on the QALYs gained. However, in practice this evidence is not available due to the multitude of interventions and the scarcity of data. Instead the comparison is often between a new technology x and current clinical practice c. The key questions for cost-effectiveness are then:

1. How much more does x cost (ΔC_x)?
2. How many more QALYs does x provide (ΔQ_x)?

Decision-making is then informed by the *cost per QALY gained*, which is often termed as the incremental cost-effectiveness ratio (ICER):

$$\text{ICER} = \frac{\text{Incremental costs of } x}{\text{Incremental QALYs for } x} = \frac{\Delta C_x}{\Delta Q_x} \tag{1.1}$$

In England, the National Institute for Health and Care Excellence (NICE) runs a (health) technology appraisal (TA) process as part of its HTA programme. For this, most health technologies with an ICER less than £20,000 are likely to be approved, whilst those with an ICER greater than £30,000 are unlikely to be approved. Between these amounts, decisions are influenced by other factors, such as innovation, that may not be quantified by the use of an ICER for decision-making. For technologies which meet certain end-of-life criteria, a higher threshold (or 'willingness to pay') of £50,000 is used, such that technologies with an ICER less than £50,000 are likely to be approved [10]. Treatments for very rare conditions are classified by NICE as highly specialised technologies. For these technologies, a default threshold of £100,000 is assumed, with higher thresholds applicable for technologies that provide significant QALY gains [11]. The English NHS is legally obliged to fund and resource health technologies that have been recommended by the NICE TA process within a certain time frame, if both the clinician and the patient think the intervention is appropriate [12]. The use of time series models for predicting future survival in a HTA is presented in Chapter 6, which uses an existing NICE TA as a case-study.

Decisions on if a health technology should be approved, and so made available to patients, depend on estimates of the costs and QALYs accrued for the technology. These in turn are influenced by estimates of survival. As detailed in the following section, the available evidence on survival is typically shorter than the required survival evidence, necessitating extrapolation.

1.2 Extrapolation of survival data

The impact of health technologies on survival, or mortality, can vary markedly over time. For example, endovascular repair of abdominal aortic aneurysms was associated with a reduced mortality rate after 30 days (compared with open repair) but an increased mortality rate after 15 years [13]. Conversely, screening for ovarian cancer had no impact on disease-specific mortality for the first seven years but reduced disease-specific mortality (compared to no screening) for the subsequent seven years [14]. These two examples demonstrate that evidence on effectiveness will depend on the time horizon used. Hence to enable a fair comparison across different health technologies, the time horizon used in HTA needs to be long-enough to include all important differences in treatment outcomes and costs [12]. If the health technology impacts on survival, then by definition a lifetime horizon is required.

However, there is typically discordance between the required evidence and the available evidence, which requires the use of extrapolation. For example, a review of oncology submissions to the NICE TA process found that the average time horizon used was 25.2 years whilst the available evidence covered an average of 2.9 years, with extrapolation required in 28 of the 29 appraisals [15]. A more recent review of oncology submissions to the NICE TA programme in 2017 found that extrapolation was required in all 28 appraisals, with the required time horizon being on average ten times longer than the maximum available follow-up (31.4 years vs 3.1 years) [16].

1.2.1 Limitations with current extrapolation methods in HTA

Survival models have been developed to describe observed survival data [17]. Current practice in HTA is to fit survival models to the observed data and use these to generate extrapolations. More details on current practice are provided in the review contained in Chapter 2 whilst more details on the technical aspects of survival data are provided in Appendix 1.

There is neither a 'gold standard' method for producing extrapolations in HTA, nor is there agreement on which models or methods should be used [18, 19]. In 2013 the NICE Decision Support Unit (DSU), which provides support to the NICE TA programme, published a Technical Support Document (TSD) covering extrapolation of survival data [20]. This document included a review of English oncology TAs completed by December 2009. The author concluded that inconsistent, and sometimes arbitrary, approaches were taken to model fitting, with flexible modelling methods rarely considered and the chosen approach never fully justified. The TSD provided practical guidance to aid in the choice of methods for analysing and extrapolating survival data. Six years later, an updated review of NICE TAs found that there was little improvement in analyses [21]. This updated review also noted that some appraisals included more complex survival models than those covered by the TSD, suggesting the need for updated guidance.

To compound the issue of inconsistent approaches to extrapolation in HTA, a number of studies have noted that competing statistical models can provide very similar fits to the observed data but markedly different extrapolations. Guyot and colleagues compared fifteen different methods to estimate the survival benefit of a cancer treatment [22]. Two methods provided almost identical fits to the observed data but very different extrapolations, with estimated survival benefits of 13.9 and 80.4 months. Estimates for the other methods ranged from 4.7 months (the preferred approach) to 195.5 months, demonstrating that goodness of fit to the observed data cannot on its own be used to choose between competing models when extrapolations are required. Similarly, estimates of cost-effectiveness, and hence funding decisions, have been observed to be sensitive to the extrapolation method used in a number of studies [6, 7, 23, 24].

The current practice of using survival models for extrapolation has also resulted in problems with how the observed data are used to inform extrapolations. Survival models (and hence their extrapolations) give more weight to times with larger sample sizes, which occurs at the start of follow-up. However, recent outcomes (at the end of follow-up) may be more informative of the future than outcomes at the start [25, 19]. The survival models used as current practice cannot accommodate this, which has motivated the use of 'piecewise' models [26]. These apply different models at different time-points. Models may be 'hybrid', using non-parametric methods initially, with a parametric model for the subsequent years [19]. Future predictions are based on estimates from the model fitted to the most recent observations (the extrapolating model). This approach has two key limitations [27, 28]:

1. The choice of which data to include within the extrapolating model is essentially arbitrary and extrapolations can be sensitive to this choice
2. Compared with using all of the data, the extrapolating model will be fitted to a reduced sample size. Often the reduction in sample size will be dramatic, which can lead to poor extrapolations with increased uncertainty.

Survival models were developed to describe the observed survival data but they were not explicitly designed to predict the future. Their use to generate extrapolations has led to the limitations described above. In contrast, generating forecasts is a main task of time series models, which avoid the aforementioned limitations. These are described next.

1.2.2 Time series models for extrapolation

The task of predicting the future is not unique to HTA. It occurs in a number of alternative disciplines including demography [29, 30], epidemiology [31], finance [32, 33], healthcare [34] information and communications technology [35] and ecology [36]. A prominent approach for generating extrapolations in these other areas is with time series models [37]. Time series models account for the fact that outcomes are correlated across time. Because they can exploit the temporal dependence amongst outcomes and are often used to predict the future in many areas outside of HTA, the motivating hypothesis of this thesis is that time series models will be of value for extrapolating survival data in HTA.

A large appeal of time series models is that they can exploit the temporal structure of the data; resulting models are known as 'local' models. For a local model, the model parameters can vary continuously over time. In contrast, survival models currently used in HTA are 'global' models. Global models have the property that the model parameters are fixed, so that they are the same at all time-points. A third type of model is a piecewise model; for this the full-time span is split into mutually exclusive (and exhaustive) subsets. Model parameters are fixed within a subset but can vary across subsets. An example of a piecewise model would be if the observed data were split into two halves, with a different model fit to each half. Global models may be viewed as a special case of both piecewise models (where there is only one subset) and local models (where model parameters do not vary). Further, for a piecewise model the full-time span may be split so that each subset contains a single event time. This piecewise model would then use the same intervals as a local model but it is not the same as a local model in general. This special case of the piecewise model is referred to as a *local piecewise model.* For this model, an independent model is fit to each time interval meaning that model parameters are independent across time and the temporal structure of the data is ignored. In contrast, a local (non-piecewise) model introduces correlations in the parameter estimates across time intervals. These temporal correlations (also known as autocorrelations) mean that the evolution of parameter estimates over time follows a time series. The use of different time series will lead to different local models. To avoid a loss of information, in this thesis the number of time intervals shall equal the number of unique event times, as recommended in the literature [38]. Previous research has suggested that results are generally robust to the choice of time intervals, provided that there is at least one event per interval (as occurs in the approach used here) [39, 40].

To highlight the differences between global and non-global models, a simple example shall be used. This uses a subset of the survival data from the case-study of Chapter 6 (restricted to less than 10 months follow-up). This data is provided in Table 1.1 and is summarised by hazard values over time (denoted by λ_{t_i}). The hazard may be viewed as the probability that a person will die during a period (interval) of time, given that they were alive at the start of that interval.

The data displayed in Table 1.1 were obtained by digitising a published Kaplan-Meier graph so there are multiple events (deaths) per unique event time. The analyses of Chapters 4 and 5 demonstrate analyses with individual patient-level data so have one event per unique event time.

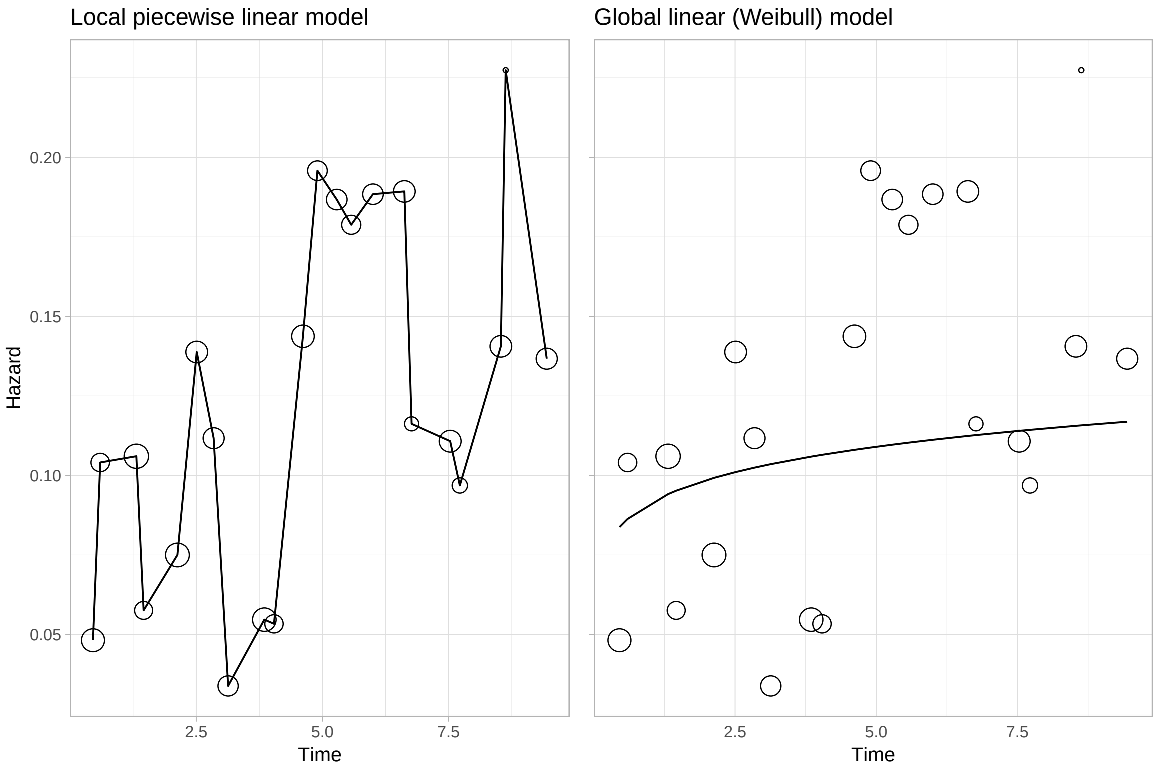
A review of survival models used for extrapolation in HTA identified the Weibull as the most frequently used model [18]. This is a linear model of the log-hazard (see my fellowship publication [1], which is reproduced in Appendix 3, for further details). As such, it has two parameters: an intercept (β_1) and a slope, or trend (β_2). In contrast, a piecewise local model divides the observed data into mutually exclusive time intervals and fits a different linear model to each interval, so the number of parameters will increase as the number of observed events increase. For local models, including a correlation between the parameters means that they are no longer independent. As such, there is no straightforward interpretation of the number of parameters in a local model.

The data of Table 1.1 relate to 22 unique death times. One approach is to fit a local piecewise model, assuming a linear change between observations. For this local piecewise linear model there is a single intercept and 22 trends (β_{2,t_i}). Estimates from this model are provided in the left-hand pane of Figure 1.1, whilst estimates from the Weibull model are displayed in the right-hand pane. The estimates from the two models are very different: the piecewise model exactly matches the data. Hence it provides a very good fit to the observations but there is a strong possibility that it is over-fitting the data, as some of the variation will be due to random noise. In contrast, there is little danger of the Weibull model over-fitting any noise in the data. It captures the general trend of the hazard to increase over time. However, there is a possibility that the Weibull model is under-fitting the data, so potentially failing to capture any important local variations about the increasing in the trend. Use of the global Weibull model also assumes that the observed trend is the same at all time-points. Use of local time series models are designed to overcome the limitations with both piecewise and global models and so identify the optimal middle-ground between over-fitting (as may occur with a piecewise model) and under-fitting (as may occur with a global model). The analysis of Table 1.1 with a local model is presented in Chapter 6, the results of which suggest that there is important local variation in the hazard estimates, but that this variation is not as extreme as implied by a local piecewise model.

Table 1.1: Data used in Figure 1.1

Time	**Alive**	**At risk**	**Events**	**Hazard**
0.45	137	62.23	3	0.05
0.60	134	19.22	2	0.10
1.31	132	94.30	10	0.11
1.46	121	17.36	1	0.06
2.13	120	79.98	6	0.08
2.51	113	43.22	6	0.14
2.84	107	35.80	4	0.11
3.13	103	29.55	1	0.03
3.85	102	73.14	4	0.05
4.04	98	18.74	1	0.05
4.61	97	55.65	8	0.14
4.90	89	25.53	5	0.20
5.28	84	32.13	6	0.19
5.57	78	22.37	4	0.18
6.00	74	31.84	6	0.19
6.62	68	42.26	8	0.19
6.76	60	8.60	1	0.12
7.53	59	45.14	5	0.11
7.72	54	10.32	1	0.10
8.53	53	42.67	6	0.14
8.63	46	4.40	1	0.23
9.44	45	36.57	5	0.14

Figure 1.1: Comparison of local and global piecewise models



As mentioned, the global model and the local piecewise model provide very different estimates in Figure 1.1. These differences are driven by the use of different structural assumptions. These structural assumptions highlight both the strengths and limitations of the two modelling approaches:

Global linear model (Weibull): This assumes that the evolution of the log-hazard function over time may be described by a single linear model with two parameters. The two main strengths of this assumption are:

- It provides a parsimonious description of the data, so is unlikely to over-fit the data.
- Estimates of the two parameters are based on all of the observed data, so will be more precise than estimates based on subsets of the data.

The main limitations with the assumption of a global linear model are:

- The assumption of global linearity is a strong one; the model will not give a good fit for non-linear hazard patterns.
- When estimating the model parameters, the most weight is given to data at the start of follow-up and the least weight is given to data at the end of follow-up (as these have the largest and smallest sample sizes, respectively). This is problematic when generating extrapolations as data at the end of follow-up are likely to be more representative of the future than data at the start of follow-up.

Local piecewise linear model: This assumes that the log-hazard function may be adequately modelled by the observed hazard rates, with a linear change in-between observations. It is able to remove the limitations of the Weibull model:

- The assumption of piecewise linearity allows this model to fit closely any type of complex hazard pattern.
- To generate extrapolations the last estimate of trend may be used. This ensures that extrapolations are based on the most recent (and hence potentially most relevant) data.

However, the piecewise linear model loses the strengths of the Weibull model:

- The observed data is likely to include some random variation; by using the observations directly the piecewise model will be over-fitting the data. This may also be noted as the model has more parameters than there are unique event times.

- Estimates of the trend at any given time are based on two observations, so will be highly variable. This can include 'jumps' in the model estimates. It will also induce additional uncertainty in extrapolations when compared to using the whole data to inform extrapolations, whilst the reduced sample size means that extrapolations may be less accurate.

- Use of different time intervals can lead to markedly different extrapolations [27] and there is no guidance on what time intervals should be used in a piecewise model.

Of note, the local piecewise linear model is a parametric model (having $d+1$ parameters, where d is the number of unique death times). It is also the same as the Kaplan-Meier estimate of the hazard, which is referred to as a non-parametric method [17]. Hence for this thesis the distinction between parametric and non-parametric methods (or models) is not important. The key distinction is the assumptions made by the differing models along with how realistic they are [41].

Piecewise models are local models but they do not exploit the temporal dependence amongst outcomes. Time series models exploit this temporal structure; doing so allows them to retain the advantages of piecewise models, whilst ameliorating their limitations. The time series models used in this thesis are an extension of *exponential smoothing* models [42]. They may be viewed as piecewise models for which parameter estimates at time t_i are correlated with previous parameter estimates (for times $< t_i$). For a local piecewise linear model the estimate of $\beta_{2,t_i} = \lambda_{t_i} - \lambda_{t_{i-1}}$ (or the equivalent on the log-scale), for an exponential smoothing model it is a weighted average of the previous model estimate and the observed data: $\beta_{2,t_i} = \alpha\beta_{2,t_{i-1}} + (1-\alpha)(\lambda_{t_i} - \lambda_{t_{i-1}})$. The advantages of using a time series local model are:

- It makes a very weak structural assumption: that data which are close together in time are similar (correlated). This allows it to fit many complex hazard patterns.

- Unlike a global model, more weight is given to more recent data when generating extrapolations. Unlike a piecewise model, all of the observed data are used when generating extrapolations (the amount of weight given to older observations decays exponentially).

- The degree of correlation between observations is estimated during model fitting and can be constrained to reduce the danger of over-fitting.
- The global and local piecewise local models are included as special cases (with correlations of one and zero, respectively).

As such, the use of time series models for generating extrapolations in HTA has strong theoretical support. In particular, their use would immediately resolve an on-going area of disagreement in the HTA community. This has seen rise to two schools of thought about the best approach to extrapolation when using common practice survival models. One is that all of the data should be used to avoid a loss of information, the other that models should be fit to a subset of the more recent data to give additional weight to the more recent observations [18, 19, 28]. There are to-date no practical applications or evaluations of the use of time series methods for extrapolation with HTA. This thesis aims to fill this evidence gap.

1.3 Aims and thesis layout

The main research question to be answered by this fellowship is: “What is the value and role of time series methods for generating predictions of future survival in HTAs?” To answer this question it is important to identify alternative methods that may also be used to predict future survival. This fellowship has the following aims:

1. Identify methods that may be used in addition to time series methods for predicting the future within the context of HTA.
2. Compare the extrapolation performance of time series and comparator methods.
3. Establish the suitability of using time series methods for extrapolation in a HTA context.
4. Develop good practice for using time series methods to generate predictions.
5. Identify areas for future research.

To achieve these aims, the following work was undertaken:

1. Establish current practice for extrapolation of survival data via a review of the methods used in English HTAs.
2. Conduct a literature review to identify alternative methods (to time series models and current practice) for the extrapolation of survival data.
3. Perform a literature review to describe the current state of the art for applying time series models to survival data. If required, develop extensions to these models in response to the challenges of extrapolation in HTA.
4. Implement simulation studies and a case-study to evaluate the extrapolation performance of the identified methods.
5. Replicate an existing HTA to demonstrate the feasibility of incorporating the output of time series models in cost-effectiveness models.
6. Use the results of the above analyses to produce good-practice guidelines.

The consideration of uncertainty in extrapolations is also very important. This uncertainty may be quantified within an economic evaluation via probabilistic sensitivity analyses and value-of-information analyses [43, 44, 45]. A full treatment of these topics is beyond the scope of this thesis so is not pursued in detail here, although it is illustrated in Section 6.4.2. Further, I have provided some exploratory work into this as part of my fellowship. This work highlighted that the choice of survival model for extrapolation can significantly impact on both the extrapolated hazard function and estimates of cost-effectiveness. These differences were observed even when the fits of different models to the observed data were similar and in general this structural uncertainty was not reflected by estimates of uncertainty from any given survival model. This work has been published as a peer-reviewed manuscript [2].

The remainder of this these is arranged as follows. Alternative extrapolation approaches to time series methods are identified in Chapter 2. This chapter includes both a review of the methodological literature and a review of NICE TAs (Sections 2.1.1 and 2.1.2 respectively). Together, these reviews help to delineate comparator methods, including both current practice approaches and emerging approaches. A framework that allows for the use of time series methods to extrapolate survival data is described in Chapter 3. This chapter also provides a focused review of relevant time series literature in Section 3.2 to identify state of the art methods. The results of this review motivated novel work that was developed as part of this thesis. This work includes extensions to the framework to make it more useful in a HTA context (Section 3.3) and a simulation study which was performed to inform model specification (Section 3.4).

Extrapolation performance was assessed using a real patient-level dataset in Section 4.1. For this case-study both time series methods and the identified competing approaches were fit to an interim analysis of trial data and extrapolations were compared against the available longer-term data. Simulation studies were also used to compare the extrapolation performance and technical properties of the available methods. These simulation studies used complex hazard patterns. In both situations the overall hazard arises as a mixture of two sub-group hazard functions. In Chapter 4 these two sub-groups represent different levels of frailty (likelihood of death over time). The situation where one group may be viewed as cured is covered by Chapter 5, which considers extrapolating models with and without a cure fraction.

As many of the statistical models considered are more complex than those routinely used in HTA, it is important to consider the feasibility of incorporating them within appraisals. This is provided in Chapter 6, which recreates an existing NICE appraisal where differing approaches to extrapolation had a substantial impact on cost-effectiveness. These approaches to extrapolation were identified as a key source of decision-making uncertainty. This chapter also illustrates the incorporation of external evidence to avoid implausible extrapolations, because this issue was highlighted in the original appraisal.

An overview of the work performed, the contribution to the knowledge base, suggested good practice guidance, and a discussion of potential avenues for future research are provided in the concluding Chapter 7.

1.4 Conclusions

Finite health care resources have led to the development of HTA as a framework for helping to inform decisions about which technologies to fund. To make fair comparisons across different technologies it is important that a common metric is used and that all differences in outcomes are incorporated within the assessment. To achieve this typically requires evidence on outcomes for a longer time-period than is available. Hence statistical methods are required to predict future outcomes based on the available evidence. The use of different statistical methods can lead to substantively different extrapolations, which in turn can lead to different funding decisions. Recent reviews have highlighted that current approaches to extrapolation in HTA are often arbitrary, with little justification and little consideration of more flexible approaches. In addition, time series methods are routinely used for making predictions in subject areas outside of HTA but to this author's knowledge have not been used in HTA. This suggests that further research is required into establishing the role and value of more flexible methods for producing extrapolations in HTA. This thesis seeks to contribute to the knowledge base and help to close this evidence gap. One of the key contributions is the introduction and evaluation of the role of time series methods when extrapolating survival data in HTA. This includes novel developments both to incorporate external evidence and also to facilitate extrapolations. A further key contribution is the rigorous and comprehensive identification and evaluation of methods for the analysis and extrapolation of survival data in a number of different situations, including curative treatments. The choice of methods to evaluate, in addition to time series methods, was informed by comprehensive reviews of both the methodological literature and the methods used in NICE TAs. These reviews and their results are detailed in the next chapter.

Chapter 2

Literature review of survival analysis methods

The primary aim of this thesis is to identify the value and role of using time series methods to analyse and extrapolate survival data as opposed to alternative methods. To achieve this aim it is important to identify the alternative methods that are available, and what should be included as comparator methods in this thesis. Two types of comparator are of interest:

Current practice. This represents statistical models that are routinely used for extrapolating survival data in HTA. As the field of HTA is broad, this review focuses on submissions to the NICE TA programme, which has long-been viewed as an internationally influential example of HTA [46].

Non-standard models. These models are defined as those that may be used to extrapolate survival data but are neither time series models nor current practice. This literature review identified more non-standard models than could be evaluated within the timescales of this fellowship. As such, a clear justification is provided for which non-standard models were retained. This justification includes a consideration of ease of use and how extrapolations are generated (for example, if the models are global or local).

The identification of these two types of comparator required two types of literature review. A review of the methodological literature was conducted to identify the available statistical methods. A review of NICE TA submissions was then performed to identify which of the available methods represented current practice. This chapter describes the process of performing these reviews, the identified models, and details on the models that shall be considered in subsequent chapters. The methods used are described in Sections 2.1.1 and 2.1.2 for the methodological review and review of NICE appraisals, respectively. The primary aim of the reviews is to identify comparator methods to time series methods, but the reviews are broad enough to also identify relevant studies that used time series methods. Results of both reviews are presented in Section 2.2. A broad overview of the resulting long-list of methods is described in Section 2.2.1. This overview provides, for each of the identified methods: the applications that it has been developed for; the strengths and limitations of this method; its use within NICE TAs; and the implications for this thesis if the method is pursued (including risks and evidence gaps). The results from this broad overview were then used to justify the short-list of methods to take forwards for further research, as discussed in Section 2.2.2. For this short-list of methods, a more detailed technical discussion including their assumptions, strengths and weaknesses is provided in the published manuscript "Generalized Linear Models (GLMs) for Flexible Parametric Modeling of the Hazard Function" (hereafter "the GLM manuscript"). This manuscript is an output from this fellowship and is reproduced as Appendix 3.

2.1 Methods

2.1.1 Methodological review

The aim of the methodological review was to identify the breadth of statistical methods available for the analysis and extrapolation of survival data. There are a wide variety of different approaches to performing literature reviews; with at least 14 main types [47]. A mapping review was chosen in consultation with a systematic reviewer at ScHARR, as it was felt that this would be the most appropriate approach. Mapping reviews attempt to categorise the available evidence and current knowledge pertaining to a broad topic of interest [48, 49]. Studies are synthesised by categorising into similar themes (or knowledge clusters), without formal quality assessment [48, 50]. The mapping review had the following two steps:

1. Conduct a mapping review of the methodological literature.
2. Use the results of the mapping review to characterise the strengths and weaknesses of the identified methods.

The mapping review took the form of comprehensive literature searches in two databases, supplemented by pearl-growing techniques [51], hand searching of both key journals and the grey literature, and a keyword search.

As the review was of the methodological literature, traditional reviewing methods required modification [52]. In particular, it was important to distinguish between two types of literature [52, 53]:

1. The methods literature, which presents a method and describes the theory underlying it, and
2. the 'case-studies' (operationalisation) literature, which provides empirical evidence about how a method may be applied in practice

The focus of the mapping review was on the methods literature. In contrast to traditional reviews, there is no need to identify every study that describes a given method as there is unlikely to be any added value from repeatedly reviewing the same methodological approach. As an example, the semi-parametric Cox proportional hazards model is a method for analysing survival data. The paper that introduced this model has been cited almost 35,000 times [54]. Hence, any search strategy which did not make a distinction between methods literature and case-studies literature, and which also sought to identify all of the articles which used statistical methods to analyse survival data is likely to identify an impractical number of articles for reviewing.

Another feature of methodological reviews is that searches should cover a broad range of evidence sources, including non-journal articles [52, 53]. If key journals are identified, hand-searching these may be a viable search strategy [53, 55]. The reviews may also be iterative, as the understanding of the available research methods (and what should be included in search terms) evolves [56, 57, 58]. More details on the search strategy, including the search terms used, are provided in Appendix 2.

The following sources were used:

- Bibliographic databases: PubMed and Web of Science (WoS). Search terms were created in consultation with an information specialist. Two separate searches were conducted. The first was to identify studies describing methods specifically for the extrapolation of survival data. The second was to identify review articles on the analysis of survival data.
- Hand-searching journal articles: Journal of the Royal Statistical Society, Series B (JRSS B): all articles. Statistics in Medicine: tutorials in biostatistics, Medical Decision Making: special issue on "Methods for Extrapolating Survival in Cost-Effectiveness Analyses". The choice of journals and articles to hand-search was a pragmatic one aimed at identifying with high specificity (precision) relevant articles, whilst also keeping the number of articles to sift at a feasible amount.
- Citation searching of key studies identified via the above methods.
- Grey literature: Academic textbooks, and the methods available in the statistical software packages STATA (user packages) and R (as listed in the CRAN Task View for survival analysis [59]).
- Key-word searching: The term "Expectation of life" was identified during the searching process (and was not part of the search terms developed for the bibliographic databases), so a keyword search for this was performed in PubMed.

For a mapping review results are categorised into themes. The definitions of these themes evolved as the understanding of both the methods and the implications of the categorisations evolved. As such, categorisation was an iterative process relying on subjective interpretation. This approach is also referred to as the concept of literary warrant [47, 60]. The subsequent classification of themes into five broad groupings likewise evolved over time. It is noted that these types of classification are by their nature subjective, and that the important aspect is that the process is explicit and transparent [61, 62]; use of a mapping review ensures that the search process is systematic.

Inclusion and exclusion criteria

As the mapping review sought to answer a broad question, the inclusion criteria were intentionally kept broad. The literature reviews included grey literature in addition to journal articles; these are collectively referred to as 'sources'. The inclusion criteria were:

1. Source contains one or more descriptions of a modelling method for the analysis of survival data.
2. English language publication.
3. (Books only): Not superseded by a more recent publication.

Further to the first inclusion criteria, sources that only included a general discussion (either of survival modeling or of extrapolation) were excluded. There were also some pragmatic exclusion criteria, as it was decided that certain methods would be outside the scope of the fellowship. The exclusion criteria for the comprehensive review were:

1. Source only describes methods for treatment switching.
2. Source only describes methods for network meta-analysis.
3. Source only describes methods for high-dimensional (or multivariate) outcomes, defined for this search as more than two outcomes.

The grey literature searches and hand searching of journals had date-based exclusions, as described in Appendix 2. All searches were performed on 31st July 2017.

2.1.2 Review of NICE appraisals

The aim of this review was to identify the type and frequency of methods employed to analyse and extrapolate survival data in practice, using the NICE TA programme. A benefit of using the NICE TAs programme as the sampling frame for identifying current practice is that it provides a very focused search, as all of the studies are in the context of HTA, and it is anticipated that the majority of the studies will include an analysis of survival data.

A subset of all NICE appraisals that had been published or updated since the start of September 2013 and the end of June 2017 was reviewed. The start date was chosen as being six months after the last update of the NICE DSU TSD on extrapolating survival data [20]. This guidance is expected to influence the approach taken to extrapolation in NICE appraisals, which were assumed to take about six months to complete. In total, there were 140 appraisals in this period. The number of appraisals to review was not pre-defined; instead reviewing occurred until it was felt that data saturation had occurred; that is further reviews would not substantially alter the results of this review. There is an element of subjectivity in identifying the saturation point, as the possibility cannot be disregarded that important methods were covered in the excluded appraisals. However, this approach maximises the number of appraisals that can be included within a time-constrained review. The appraisals were reviewed in a randomly determined order. Whilst conducting this review, an independent review of NICE cancer appraisals was started in ScHARR (commissioned by Bristol-Myers Squibb, hereafter referred to as the 'BMS review'). This reviewed all cancer appraisals between the start of July 2011 (the initial publication of the DSU TSD on extrapolation), and the end of June 2017. For my review I included 62 appraisals, of which 32 were cancer appraisals. The BMS review included a further 30 appraisals that I did not review. As the results from the BMS reviews are relevant to this work, they are discussed jointly here. Hence, 92 appraisals are included in this review, of which 62 were reviewed by myself (67%). These reviews cover all cancer appraisals that commenced since the publication of the DSU TSD on survival modelling [20], along with a simple random sample of non-cancer appraisals. Detailed results for the BMS review are available in an open-access publication [21] (there is a minor difference in the included cancer appraisals, as I included multiple technology appraisals, whilst the BMS review excluded these).

For each appraisal, details were extracted on the approach described within the company (or analysis group) submission for analysing and extrapolating survival data (when applicable). If there were multiple survival outcomes, evidence was extracted for the outcome which had the most detailed description of the modelling approach. Two previous reviews of NICE TAs have focused exclusively on cancer submissions ([18, 63]), so results were stratified by whether or not the TA was cancer-related.

2.2 Results

For the methodological mapping review, the number of results by evidence source is provided in Table 2.1. Figures 2.1, 2.2 and provide modified PRISMA diagrams for the mapping review, providing details on the search process by evidence source.

Table 2.1: Results by evidence source

Source	**Identified**	**Retained**
Web of Science: extrapolation studies	526	11
PubMed: extrapolation studies	807	17
Hand-searching Medical Decision Making	8	8
“Expectation of life” keyword search	320	0
PubMed: statistical reviews	334	10
PubMed: HTA reviews	1106	7
Citation searching	113	5
Hand-searching Statistics in Medicine	79	8
Hand-searching JRSS B	287	10
Academic textbooks	21	10
Software packages	260	32
Total	**3861**	**118**

The review of extrapolation studies returned 1,333 articles (807 from PubMed and 526 from WoS). After removing duplicates, 1,311 articles remained. Of these, 18 relevant articles were identified (10 were identified by both sources: [18, 19, 26, 64, 65, 66, 67, 68, 69, 70], seven by PubMed alone: [23, 71, 72, 73, 74, 75, 76] and one by WoS alone: [77]). The Medical Decision Making special issue provided an additional eight articles [22, 78, 79, 80, 81, 82, 83, 84]. The targeted keyword search for “Expectation of life” did not result in any additional studies.

The review of statistical reviews identified 10 relevant articles [18, 85, 86, 87, 88, 89, 90, 91, 92, 93]. The review of HTA reviews identified 6 articles [18, 19, 55, 94, 95, 96], in addition to one article from the grey literature that was already known [15]. Five of the articles only considered standard survival models with at most two parameters. No review sought to identify all of the available methods for the analysis and extrapolation of survival data, nor were time series methods considered.

Citation searching led to a total of 113 studies, of which eight were highlighted as being relevant. Two of these were existing pearls [18, 19], and one provided an overview of an existing pearl

[55], hence there were five additional studies [97, 98, 99, 100, 101]. All of these cited the pearl by Latimer [18], which also had the highest number of citations (42). One study also cited the pearl by Guyot [94].

Hand-searching journals resulted in the identification of eight tutorials from Statistics in Medicine [76, 102, 103, 104, 105, 106, 107, 108] and 10 articles from JRSS Series B [109, 110, 111, 112, 113, 114, 115, 116, 117, 118]. The grey literature identified 10 academic textbooks that were of relevance [17, 119, 120, 121, 122, 123, 124, 125, 126, 127], along with the methods used by R and STATA.

Figure 2.1: PRISMA flow diagram: reviews of reviews and citation searching

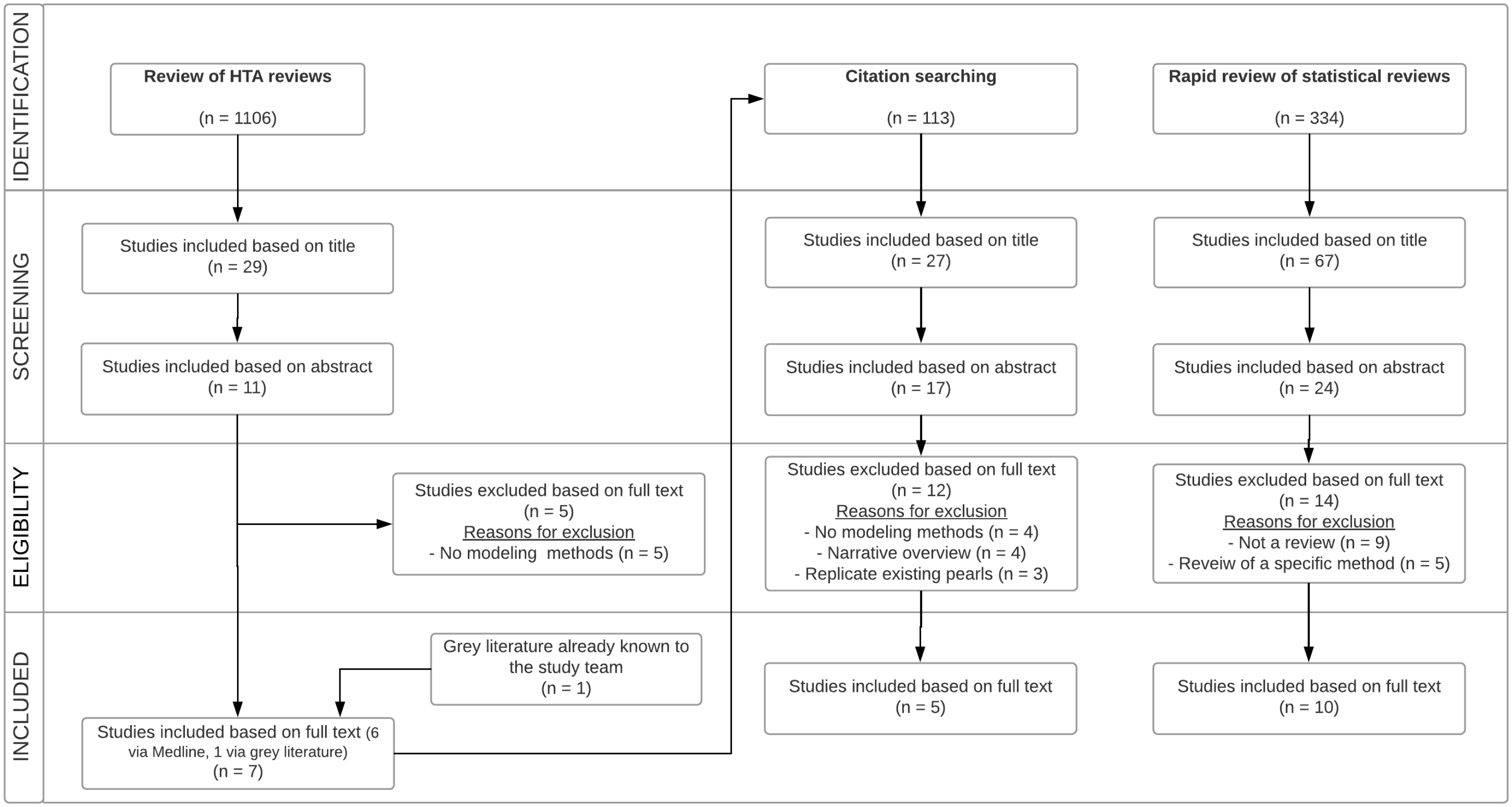

IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Review of HTA reviews
(n = 1106)
Citation searching
(n = 113)
Rapid review of statistical reviews
(n = 334)
Studies included based on title
(n = 29)
Studies included based on title
(n = 27)
Studies included based on title
(n = 67)
Studies included based on abstract
(n = 11)
Studies included based on abstract
(n = 17)
Studies included based on abstract
(n = 24)
Studies excluded based on full text
(n = 5)
Reasons for exclusion
- No modeling methods (n = 5)
Studies excluded based on full text
(n = 12)
Reasons for exclusion
- No modeling methods (n = 4)
- Narrative overview (n = 4)
- Replicate existing pearls (n = 3)
Studies excluded based on full text
(n = 14)
Reasons for exclusion
- Not a review (n = 9)
- Reveiw of a specific method (n = 5)
Grey literature already known to the study team
(n = 1)
Studies included based on full text (6 via Medline, 1 via grey literature)
(n = 7)
Studies included based on full text
(n = 5)
Studies included based on full text
(n = 10)

Figure 2.2: PRISMA flow diagram: hand-searches and grey literature

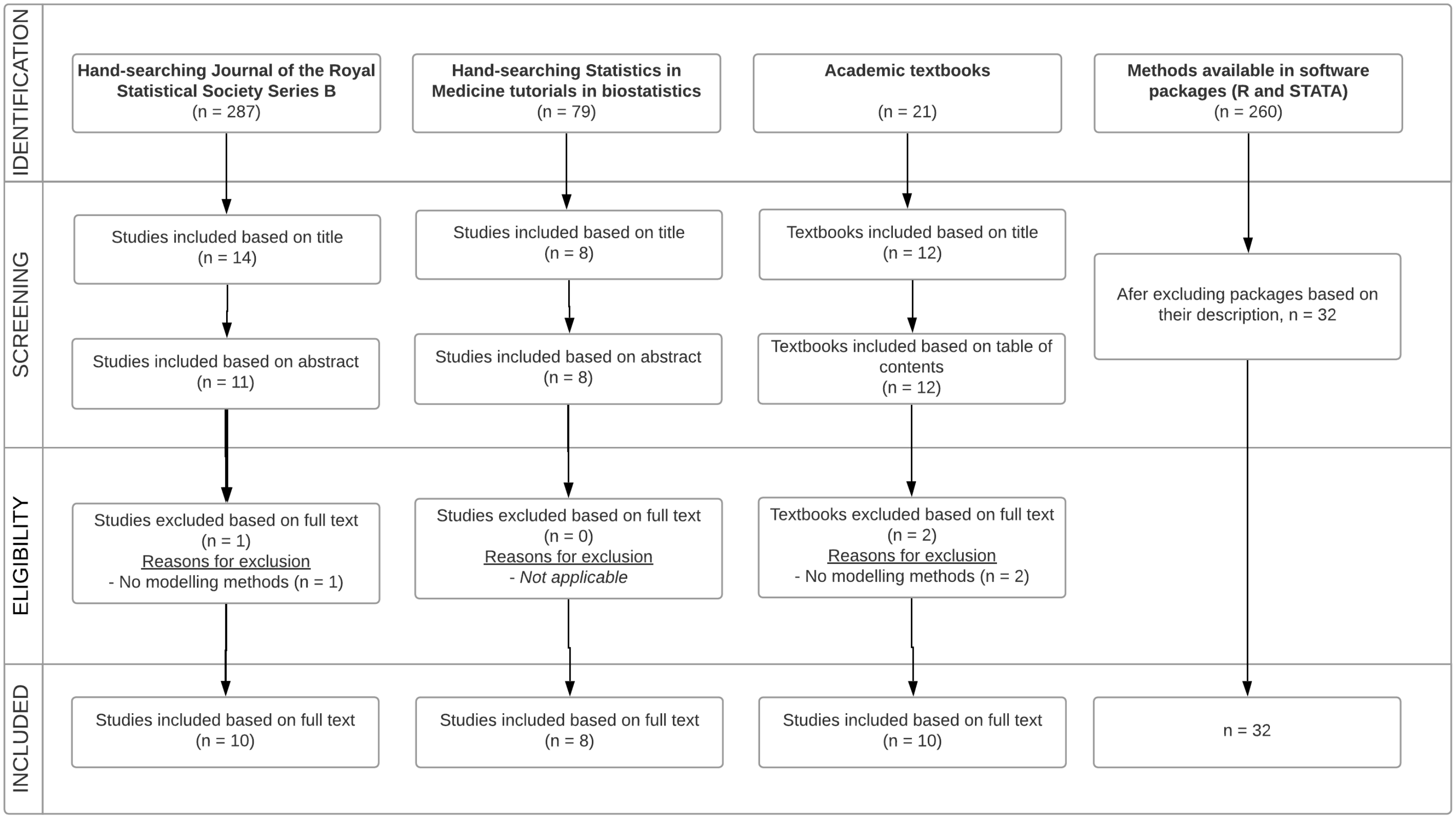
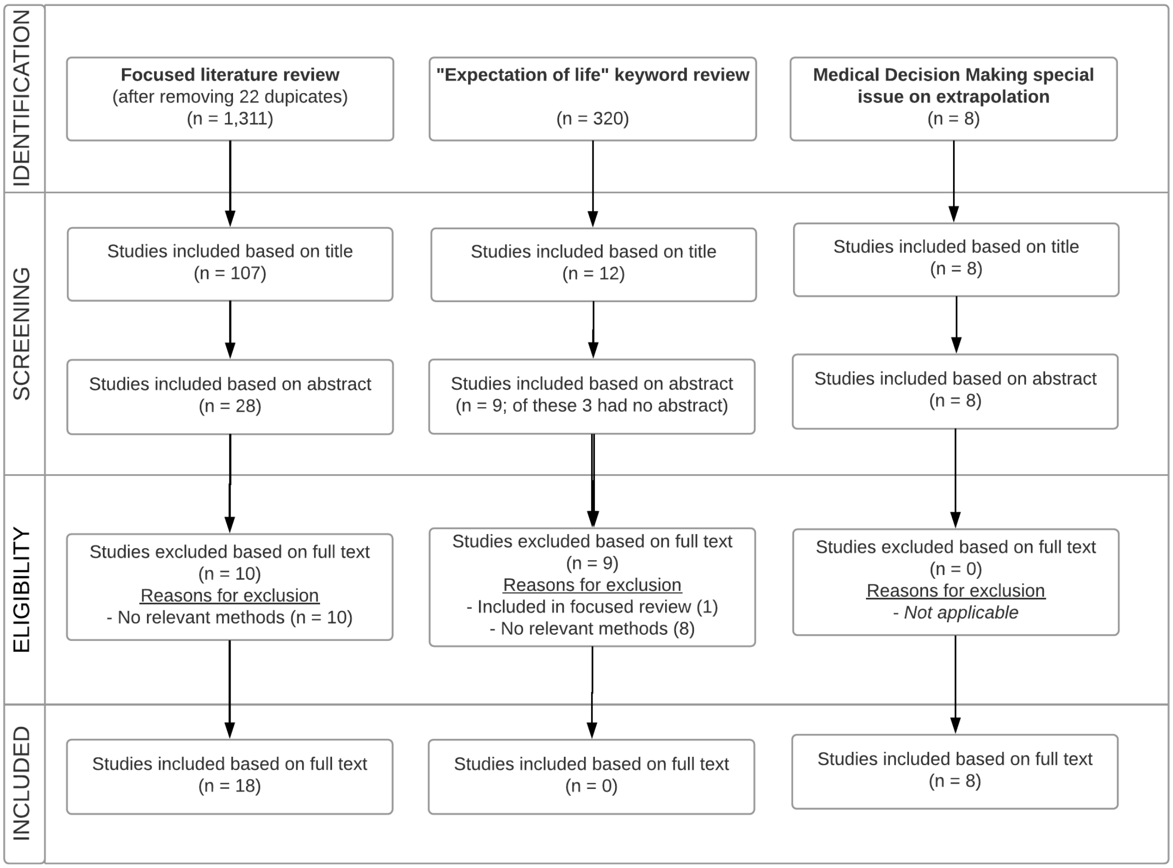

IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Hand-searching Journal of the Royal Statistical Society Series B (n = 287)
Hand-searching Statistics in Medicine tutorials in biostatistics (n = 79)
Academic textbooks (n = 21)
Methods available in software packages (R and STATA) (n = 260)
Studies included based on title (n = 14)
Studies included based on title (n = 8)
Textbooks included based on title (n = 12)
Afer excluding packages based on their description, n = 32
Studies included based on abstract (n = 11)
Studies included based on abstract (n = 8)
Textbooks included based on table of contents (n = 12)
Studies excluded based on full text (n = 1) Reasons for exclusion - No modelling methods (n = 1)
Studies excluded based on full text (n = 0) Reasons for exclusion - Not applicable
Textbooks excluded based on full text (n = 2) Reasons for exclusion - No modelling methods (n = 2)
Studies included based on full text (n = 10)
Studies included based on full text (n = 8)
Studies included based on full text (n = 10)
n = 32

Figure 2.3: PRISMA flow diagram: extrapolation specific reviews


IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Focused literature review
(after removing 22 dupicates)
(n = 1,311)
"Expectation of life" keyword review
(n = 320)
Medical Decision Making special issue on extrapolation
(n = 8)
Studies included based on title
(n = 107)
Studies included based on title
(n = 12)
Studies included based on title
(n = 8)
Studies included based on abstract
(n = 28)
Studies included based on abstract
(n = 9; of these 3 had no abstract)
Studies included based on abstract
(n = 8)
Studies excluded based on full text
(n = 10)
Reasons for exclusion
- No relevant methods (n = 10)
Studies excluded based on full text
(n = 9)
Reasons for exclusion
- Included in focused review (1)
- No relevant methods (8)
Studies excluded based on full text
(n = 0)
Reasons for exclusion
- Not applicable
Studies included based on full text
(n = 18)
Studies included based on full text
(n = 0)
Studies included based on full text
(n = 8)

Across all of the evidence sources, there were a total of 118 retained results, of which 76 were journal articles. Some of the retained journal articles were duplicated across evidence sources, removing these led to 62 unique retained articles. The identified approaches for analysing and extrapolating survival data described in all the retained results were classified into 19 methods, as detailed in Table 2.2 (along with an additional method, as described below). These were in turn grouped into five broad categories: models for which time to event is the outcome, models for which the hazard (or equivalently, the survival) function is the outcome, models that account for patient heterogeneity, models for patient histories, and an 'other' category for all other approaches. This grouping was based primarily on the type of research question that the methods aimed to answer.

The 92 NICE appraisals reported the results of 95 analyses (65 cancer). Of these, 81 analysed and extrapolated survival data (65 cancer). Table 2.2 also includes details on the percentage of NICE appraisals that included each method. All but one of the NICE appraisals that included extrapolation considered standard 1 or 2 parameter models. For the one appraisal that did not (TA414), the company applied a mixture-cure model, arguing that standard parametric models would not be appropriate. Of note, five appraisals (one cancer) did not fit an explicit parametric model, but instead applied a constant hazard over time, which is equivalent to using an exponential model (so was classified as such). The only three-parameter model encountered was the generalised gamma, which was used more frequently in cancer appraisals than non-cancer appraisals (42% vs 19%). Other notable differences were the use of hybrid models, which were used in 22% of cancer appraisals but not considered in the non-cancer appraisals. In contrast, age-time models (which included the impact of ageing on the outcome when extrapolating without using the framework of cure models, as described in more detail in Section 2.2.1) were used more commonly in non-cancer appraisals. Use of the remaining 15 methods was very low in all TAs, with 12 methods not considered in any appraisal. One method was not identified by the methodological mapping review but was used in a NICE appraisal. This was the two-parameter Extreme value model, which was used in four appraisals (three cancer). Of note, this model always provided a worse within-sample goodness of fit than all of the other two-parameter models considered (as a minimum, the Weibull, log-logistic and lognormal were always also considered). Hence this model is not considered further.

The methodological review identified one approach that used time series methods: Lee-Carter models [73]. These are categorised as ‘age-time (non-cure) models’ and are described in more detail in Appendix 2 (“Literature review: incorporating external data”). There were no time series methods used to extrapolate survival data in NICE TAs. An additional time series method is the use of dynamic models [128]. These are introduced and described in the next chapter, along with justification for why they are the primary focus of this thesis. The use of dynamic time series models for the analysis and extrapolation of survival data was not identified in any of the reviews, but is included in Table 2.2 for completeness. An overview of the strengths and limitations of both dynamic models and the identified methods is provided in the next section. For each identified method this includes a discussion of the potential implications for this thesis if it were retained as a comparator. As analyses in this thesis use R, this overview also includes a discussion of the available R packages known to the author.

Table 2.2: Long-list of identified modelling approaches from the methodological review, and their use in NICE appraisals

Method	Cancer TAs (n = 65)	Other TAs (n = 16)
Time as the outcome models		
Standard 1 and 2 parameter models	64 (98.5%)	16 (100%)
Three parameter models	27 (41.5%)	3 (18.8%)
Generalised F	0 (0%)	0 (0%)
Hybrid (Kaplan-Meier and parametric) models	14 (21.5%)	0 (0%)
Non or semi-parametric models	0 (0%)	0 (0%)
Hazard as the outcome models		
Royston-Parmar (R-P) spline models	2 (3.1%)	0 (0%)
Fractional polynomials (FPs)	0 (0%)	0 (0%)
Generalised additive models (GAMs)	0 (0%)	0 (0%)
Modelling patient heterogeneity		
Mixture (non-cure) models	0 (0%)	0 (0%)
Cure models	1 (1.5%)	0 (0%)
Polyhazard models	0 (0%)	0 (0%)
Frailty (multilevel) models	0 (0%)	0 (0%)
Modelling patient histories		
Multistate models	0 (0%)	0 (0%)
Competing risks models	0 (0%)	1 (6.3%)
Recurrent events models	0 (0%)	0 (0%)
Other modelling approaches		
Age-time (non-cure) models	3 (4.6%)	5 (31.3%)
Relative survival models	0 (0%)	0 (0%)
Joint models (survival and longitudinal)	0 (0%)	0 (0%)
Dynamic models for survival data*	0 (0%)	0 (0%)

*Not identified in any reviews but included as the primary focus of this thesis.

2.2.1 Overview of the identified methods

Time as the outcome models

Overview and research questions developed for: Time as the outcome models were developed specifically to analyse survival data. These are suitable when only one outcome can be observed for each individual, and it may only be observed once. A prominent example is overall survival. The use of parametric models to analyse survival data has a long history, with an early example provided in 1952 [129]. Over time a wide variety of such models have been proposed: some key models are described in Appendix 1. Many of the models can be viewed as special cases of the four-parameter generalised F [130]. In general, models with more parameters are more flexible. Standard one and two parameter models (e.g. the exponential and Weibull) have been used in a number of appraisals [20, 95], as have piecewise models [20, 26].

Strengths: A strength of these parametric models is that a wide variety of different model specifications are available, with more complex models allowing more flexibility. For example, there are at least 14 members of the generalised F family ([130]), and models with three or more parameters are generally flexible enough to model complex hazard functions [102, 131]. Further flexibility is afforded by the use of hybrid (or piecewise) models, which fit different models to subsets of the data, with extrapolations based on the most recent (and hence most relevant) section of the data. Semi-parametric models make fewer assumptions than fully parametric models, so are more robust to errors due to model misspecification.

Limitations: The variety of models is also a drawback as it is often not clear how model selection should be performed; hypothesis testing cannot generally be used to choose between candidate models and within-sample goodness of fit may not predict extrapolation performance. Models with few parameters may impose unrealistic assumptions, such as no turning points in the hazard function. Models with more parameters are more flexible, but require larger datasets to avoid model instability, and may lead to over-fitting. Limitations with the use of hybrid models for extrapolation include a loss of information and sensitivity to the arbitrary choice of which data to include in the extrapolating model. Both limitations will lead to additional uncertainty in extrapolations, compared with using a single model at all time-points. The use of piecewise models also leads to 'step' changes in the estimated hazard, which may not be clinically plausible [98]. To extrapolate from semi-parametric models, Bayesian methods are required with a parametric prior; the sensitivity of extrapolations to the choice of prior distribution was not assessed in the identified literature, so is unclear.

Implications for this thesis: Implementation for many of these methods in R is not an issue, as eight standard models are available within the package `flexsurv` [100]. This includes the generalised F; additional models could in theory be derived as special cases of this.

For this thesis, the main drawback of considering these models is that it is unlikely to add any value to the existing literature. The models are well described in both the HTA [19, 20, 94, 95] and the extrapolation literature [23, 26, 69, 83]. Further, except for the generalised F, many of the models have been used in a number of NICE appraisals (see Table 2.3).

Hazard as the outcome models

Overview and research questions developed for: Models in this category were not originally developed for the analysis of survival data. Instead, they were designed to provide flexible descriptions of non-censored outcomes, replacing potentially restrictive structural assumptions (such as linearity) with a less strong assumption of smoothness. They can be used to model the hazard, as described in the GLM manuscript (Appendix 3). Fractional polynomials (FPs) represent the effect of time on the hazard by a combination of polynomials (typically one or two), where the powers for the polynomials are estimated from the data and taken from the set $(-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3)$, or the logarithm of the variable is selected [132, 133].

There are different approaches to the specification of spline models. One is to split the observed follow-up time into a series of intervals and fit a separate model to each interval. If cubic polynomials are used with certain restrictions (such as not having discontinuities across intervals) this leads to restricted cubic splines (RCSs) [22, 71, 134]. An alternative uses a mixture of global splines with penalisation for complexity; for this thesis this approach is referred to as a generalised additive model (GAM) [97, 135]. The application of RCS models to survival data was first described by Royston and Parmar [136], such models are referred to as Royston-Parmar models (RPMs) in this thesis and have been used to extrapolate survival data [22, 71, 80]. Contrastingly, there were no identified examples of applying FPs or GAMs to extrapolate survival data in HTA.

Strengths: A strength of both spline-based models and fractional polynomials is their flexibility for modelling complex hazard functions. Unlike standard survival models, multiple turning points in the hazard can be captured, and the models can be made arbitrarily flexible by incorporating more parameters. These models can also be combined with some of the other identified methods (such as being used in a cure model). As such, they can be applied to most research questions.

Limitations: All of the methods can require a large number of parameters, so require large sample sizes for stable estimation and to avoid overfitting (or in the case of FPs, to detect non-linear functions), with more flexible models requiring more data. There may also be uncertainty over the best model to use. A further limitation of these methods is that as they were not developed for the analysis of survival data, difficulties may arise when applying them in this context, such as estimating negative hazards (if modelling the cumulative hazard, and it is estimated to decrease [136]).

Implications for this thesis: The lack of examples that apply FPs or GAMs to survival data represents both a risk and an opportunity for this work. It is a risk as it is unclear how feasible such an application would be (as there do not appear to be R packages available that explicitly apply FPs or GAMs to survival data). It is an opportunity, as such work would represent a useful addition to both the theoretical and applied literature for HTA extrapolation. Whilst the use of RPMs is well described in the literature, there are few examples of their use in NICE appraisals.

Modelling patient heterogeneity

Overview and research questions developed for: This category of models is useful in situations when outcomes are believed to vary because of latent (unobserved) characteristics. Mixture ([137]) and polyhazard ([67, 64]) models both assume that the unobserved variables may be represented by multiple sub-models which vary with regards to how these sub-models are derived and estimated. Sometimes it may be assumed that a sub-group will become 'cured' ([17]); methods to model this are known as cure models, which typically take the form of mixture models ([71, 75]). An alternative approach is frailty (multi-level) models ([17, 107]) which model additional variation in the outcome due to unmeasured variables, with the result that some sub-groups are more prone to the outcome (more frail) than others. This is achieved by the inclusion of random-effects. Frailty models may also be used when data are collected on multiple levels; for example clustering of individuals by centre within a multicentre trial [91].

Strengths: Mixture (including mixture cure) models may often be motivated on clinical grounds; for example if it is believed *a priori* that a certain sub-group may become cured. In addition, the explicit modelling of heterogeneity can lead to more accurate estimates of uncertainty for covariate effects. These models are generally very flexible, and so may be used for survival analysis and extrapolation, even if they are not clinically motivated [138].

Limitations: As the heterogeneity is unobserved, model specification and justification (such as the number of sub-models or frailty terms) can be difficult [107], which can lead to difficulties in model fitting [17, 99, 127]. In particular, if there are no covariates in the model, frailty models are not uniquely identifiable [139], meaning that different assumptions can lead to different results.

Implications for this thesis: There is a general awareness of models for patient heterogeneity within the HTA extrapolation literature [19, 64, 67, 75], but only one of the NICE appraisals used any of these models (a cure model). At the time of conducting this review, R packages do not appear to be available for either polyhazard or parametric mixture models. A package for mixture cure models is however available ('cuRe' [140]), and it is possible to incorporate random effect terms within the GLM survival framework (as described in the GLM manuscript of Appendix 3). As such it would be feasible to include mixture cure models or random effects within this PhD. For the other models there is a risk that a large proportion of the PhD would be devoted to correctly implementing these models, which would detract from illustrating their applied use in HTA.

Modelling patient histories (multiple outcomes)

Overview and research questions developed for: These models are appropriate for the analysis of survival data when multiple outcomes may occur, and there is interest in modelling the event histories of individuals. There are two main situations: first if a patient may experience recurring events such as repeated hospital admissions; secondly if interest is in competing, mutually exclusive, events such as cancer and non-cancer causes of death. Models are available for the analysis of recurrent events [99, 124] or competing events [127, 141] in isolation. A broader class of models, known as multistate models (MSMs), is able to model both types of patient history [17, 100]: competing risks (events) are represented by different 'states', and transitions between these are defined, which allows for recurrent events. Standard survival models may be viewed as a special case of MSMs, with two health states of 'alive' and 'dead' [100] and no recurrent events.

Strengths: Advantages of MSMs are their applicability to a broad range of different research questions, as well as their potential for flexibly modelling more than one outcome [17]. They may also be used directly in HTA; serving as both the statistical and health economic model [84].

Limitations: The main drawbacks with MSMs are that model specification can be difficult [83], and a high amount of both data and data manipulation are required for use [123].

Implications for this thesis: The use of MSMs for extrapolation in HTA has been described in a tutorial paper [84], and it has also been compared with the use of other health economic models [83]. However, this only covered the use of MSMs for competing risks. Another study described the use of competing risks models for extrapolation in a non-HTA context [81]. These three papers were the only identified papers that covered this category in either a HTA or an extrapolation context. Only one example of these models was identified in the NICE appraisal review (of a competing risks model). As such, there is a possibility for further work into the role of MSMs for recurrent events in HTA. An R package exists for MSMs [142].

Other modelling approaches

Overview and research questions developed for: These are models that do not fit within the previous categories. Joint models were designed for the analysis of both survival and longitudinal data [55, 110]. Age-time (non-cure) models were developed for extrapolation (albeit not always of survival data) and take advantage of the known future ageing of a cohort. These may be applied either assuming that age and time effects are independent [23, 76], or that they interact, using Lee-Carter models [72, 73]. Relative survival models were designed to estimate disease-specific mortality from all-cause mortality where cause of death is unknown [66, 71, 79, 124]. Dynamic time series models exploit the temporal structure of survival data, as outcomes are observed over time. The evolution of model parameters is assumed to follow a time series, therefore models are very flexible [128, 143].

Strengths: With the exception of dynamic models, the identified approaches all incorporate additional evidence (on all cause mortality or the effects of age or a longitudinal covariate), which is likely to improve accuracy, and may facilitate the use of external evidence for informing extrapolations. This may be particularly advantageous if the external evidence covers a longer time horizon than the survival data. Dynamic models only impose the weak structural assumption that estimates of the hazard function which are close together in time will be similar, with the magnitude of similarity estimated during model fitting.

Limitations: The required additional evidence may not always be available. When it is, there are often multiple approaches, with uncertainty about which should be used. This means that generating extrapolations within the context of HTA may not be straightforward. For dynamic models, their flexibility can sometimes provide difficulties in parameter estimation.

Implications for this thesis: Dynamic models are the primary focus of this thesis. Neither Lee-Carter nor relative survival models have been described in the identified HTA literature. Hence future work could explore their use in this setting. In addition, the only method that was used in the NICE appraisals was age-time models, for which all implementations assumed independent effects. This suggests that further research into implementing any of these other methods may be beneficial. R packages exist for all of the approaches, although it is unclear if these may be used to generate extrapolations for survival data (due to a lack of case-studies), so there is a risk that additional development time may be required.

2.2.2 Identifying comparators

The most commonly used models were the one, two, and three parameter 'time as the outcome' models. More details on the specific model types is provided in Table 2.3. This lists the survival models that were considered in the appraisals that included extrapolation. On average four different models were considered in each appraisal. Overall, the two most popular models were the exponential and Weibull, being considered in 91% and 80% of appraisals that included extrapolation, respectively. These two models were followed by the log-logistic and lognormal models, which were both used in three-quarters of all appraisals. The generalised gamma was the only three-parameter model used, it was often unclear if this model was used or if the two-parameter gamma model was used; results are provided for these two models combined. Table 2.3 lists seven methods that may collectively be referred to as current practice.

Table 2.3: Models commonly used in NICE appraisals (gamma is both two and three-parameter versions)

Model	**Cancer Tas (n = 65)**	**Other TAs (n = 16)**	**Total (n = 81)**
Exponential	59 (91%)	15 (94%)	74 (91%)
Weibull	54 (83%)	11 (69%)	65 (80%)
Log-logistic	51 (78%)	10 (63%)	61 (75%)
Lognormal	51 (78%)	9 (56%)	60 (74%)
Gompertz	38 (58%)	8 (50%)	46 (57%)
Gamma	27 (42%)	3 (19%)	30 (37%)

In addition to current practice, a choice is required about the models that shall be used to represent non-standard comparator models. Of the 18 potential comparator methods identified in Table 2.2, two methods are categorised as representing current practice (standard one and two parameter models, and three parameter models). It would not be feasible to include the remaining 16 methods (some of which cover multiple models or model specifications) as comparators. Instead, a pragmatic approach was to restrict comparisons to those approaches that may also be implemented within the GLM survival framework described in the GLM manuscript (Appendix 3). Dynamic time series models are also implemented in (an extension of) the GLM survival framework, hence restricting non-standard comparators to GLM-based approaches increases their comparability with dynamic models. These models were further restricted to those that could be fit in R. The resulting models are:

- First-order fractional polynomials.
- Second-order fractional polynomials.
- RPMs.
- GAMs.
- Cure models.

To summarise, in addition to current practice models there are five types of comparator model. These represent more complex ‘emerging practice’ models. The GLM manuscript provides an example of fitting four of these emerging practice models (cure models are not considered, as the case-study was not for a curative treatment). The GLM manuscript (Appendix 3) also includes extensions of the models to incorporate random effects. This extension was not considered further in this thesis for two reasons. First, at a practical level, in the GLM manuscript case-study this extension did not improve the extrapolations. Secondly, at a theoretical level it is not anticipated that including random effects would have a noticeable impact on extrapolations. This is because the random effect terms do not make any explicit assumptions about the evolution of the model parameters over time (such as global or local changes), which is in contrast to the five models chosen as emerging practice comparators. The evolution of model parameters over time for the models considered in this thesis is provided in Table 2.4. It is noted that of the emerging practice models included as comparators there were some examples of use in NICE appraisals, with two considering RPMs and one including a cure model.

Table 2.4: Evolution of model parameters over time

Method	**Modelling of the hazard function over time**
Dynamic survival models	Local (correlated changes over time)
Current practice	Global (does not change over time)
Fractional polynomials (1st and 2nd order)	Global (does not change over time)
Royston-Parmar	Piecewise (changes across intervals but not within intervals)
Generalised additive models	Global (does not change over time)
Mixture cure	Local (Weighted average of two global models, where the weights change over time)

2.3 Conclusions

The mapping review of the methodological literature identified 19 models (or classes of model) for the analysis and extrapolation of survival data, in addition to the use of dynamic models. These were in turn broadly categorised into five categories. The most prominent approaches to extrapolation in NICE appraisals were the use of standard one, two or three parameter survival models, either on their own or as part of a hybrid approach where they are combined with Kaplan-Meier estimates of the hazard for earlier time-points. Of these standard models, the most frequently considered were the exponential and Weibull. These results demonstrate little change in the approach taken to extrapolation in NICE TAs up to the end of June 2017. A review of NICE TAs between September 2004 and June 2008 found that the exponential and Weibull were the most popular options [94]; the same finding was obtained in a review of NICE cancer TAs completed by December 2009 [18]. Whilst there has been no change to the most common extrapolation approaches between the earliest review and this review, there is an emerging interest in more complex survival models such as cure models and spline-based models, as neither of these were identified in the two previous reviews.

A strength of this review is the variety of different evidence sources considered, which included relevant grey literature. There are currently, to this author's knowledge, no published methodological-focused reviews of the analysis and extrapolation of survival data. As such, the results of this review help to fill an important evidence gap by systematically and transparently identifying methods for the analysis and extrapolation of survival data. In addition, the explicit comparison of the available statistical models with those that have been operationalised in NICE appraisals helps to identify implementation gaps. Several models, such as FPs, GAMs, polyhazard models and joint models were identified that may be used for extrapolation, but to the date of this review have not been used. This suggests that there are a number of useful survival modelling approaches that could be considered in the context of extrapolation; some but not all of these approaches are considered in this thesis.

There are limitations to this review. There is currently no gold-standard method for identifying methodological articles. A comprehensive search strategy was developed. This was designed to identify the majority of available methods, but it is not guaranteed to be exhaustive. It is unclear if an exhaustive and feasible search strategy for methodological studies is possible; further research into this aspect of evidence generation would be beneficial. There are at least two examples of models that were not identified by the methodological review. The first is extreme value models, which were used in four NICE appraisals (but always gave a worse fit than the other models considered). The second example is dynamic time series models, which were not identified by either the methodological review or the review of NICE appraisals. This highlights the low awareness and use of this method in medical statistics, reinforcing the importance of work on this potentially useful survival model. Of the methods that were identified via the methodological review, it is anticipated that all the main comparators to dynamic models have been identified. A further limitation is that the extraction of evidence from NICE TAs was limited. This is because not all the required evidence was made publicly available; in particular appendices (which may contain further statistical analyses) were usually not available. In addition, it is important to distinguish between what may be used for the analysis and extrapolation of time-to-event data, and what should be used. The lack of implementation of certain models may be for a variety of reasons. For example, there may be a lack of awareness of the model, software for implementing the model may not be available, or the model may not be suitable (either for the extrapolation of data, or for use within the context of HTA). However, many of the models (such as GAMs and FPs) are general methods that are likely to be applicable in any setting. As such, it is likely that the lack of use of methods in HTA is due to a lack of uptake. This review has identified methods where further research into the barriers and facilitators of uptake would be useful.

The analysis and extrapolation of survival data is an important topic and an active area of research. It is expected that the methods used within NICE appraisals may evolve over-time, with increased use of more complex methods. For example, with an increasing prevalence of curative treatments, the use of cure models or other flexible models to capture complex hazard patterns may become more widespread. The methods identified as ‘current practice’ in this chapter serve as a useful benchmark. These current practice models have a strong history of use in appraisals and are straightforward to use in analysis. The inclusion of additional comparators in this thesis means that the results shall remain relevant even if the types of survival model commonly used in appraisals changes in the future.

In addition to dynamic time series models, seven models were identified as representing current practice (exponential, Weibull, Gompertz, log-logistic, lognormal, gamma, and generalised gamma) and five model classes were identified as representing emerging practice (first-order FPs, second-order FPs, RPMs, GAMs, and cure models). Collectively these represent the comparator methods to the use of time series methods. The next chapter provides a detailed overview of how time series methods and survival analysis methods may be combined, along with novel extensions which make this combination more useful within the context of HTA.

Chapter 3

Combining time series and survival models

When time series models are combined with survival models the resulting model is known as a *dynamic survival model* (DSM). These models were introduced in the mid 1980's but do not appear to be often used within medical studies [144]. The reviews within Chapter 2, did not identify any examples of DSMs in HTA. As DSMs represent the primary focus of this thesis, this chapter provides an overview of DSMs, including their technical specification. An alternative approach to DSMs (which amalgamate time series and survival models) would be to directly apply time series models to survival data. This chapter begins with a discussion of why time series models are not directly used, which motivates the synthesis of time series and survival models. This synthesis can be achieved using the framework of dynamic GLMs. Dynamic GLMs extend GLMs to allow model parameters to vary over time and hence represent a very flexible family of models, including DSMs.

To inform the current state of the art for DSM specification and implementation, a literature review was performed. The results of this are presented in Section 3.2. Novel extensions to the DSM were developed for the analyses presented in this thesis. The technical details of these extensions are provided in Section 3.3. In addition, a key difference between global (standard) and local (dynamic) survival models is the inclusion of a parameter known as the *innovation variance*. It is important that this parameter is appropriately included within a DSM, but there has been little research into the best methods for doing so. This is explored within a short simulation study.

3.1 Dynamic survival models

Survival data may be characterised by the number of deaths per time interval, which forms a time series. There are a number of theoretical advantages to using time series models for the extrapolation of survival data, as outlined in Chapter 1. Given these, a natural question is why are time series models not currently used in HTAs? A potential explanation is that there are three key characteristics of survival data that standard time series models cannot accommodate:

Censoring: Observed survival times are censored if the outcome event has not occurred. This may be because the person has dropped-out of the study (so is no longer followed-up), or due to incomplete follow-up of an at-risk person. As censoring is a feature specific to survival data, methods to handle censored outcomes are not routinely incorporated in time series models.

Varying sample size: In survival studies it is important to account for the set of people who are at-risk of experiencing the outcome event (such as death). By definition, after experiencing the event these people will no longer be at-risk, neither will people who are censored. Hence, due to the occurrence of both events and censoring, the available sample size will decrease over time. Typical time series problems (such as estimating the future temperature or rainfall in a region) are for tasks for which there is no well-defined sample size.

Unequally spaced observations: Time series data are usually aggregated across equal time intervals (for example, the average daily temperature or annual rainfall). It is possible to represent survival data using equally spaced intervals, but this leads to a loss of information, so unequally spaced intervals (with one event per interval) are used in this thesis.

Because of these three reasons, time series models cannot be directly combined with the standard specification of survival models (as provided in Appendix 1 where, for example, the likelihood function includes information on censoring). This includes the use of Lee-Carter models, a type of time series method identified in the comprehensive review of Chapter 2. Lee-Carter models may be used to extrapolate summary mortality data, assuming that observations are equally spaced over time, with no censoring and no differences in the underlying sample size. For the approach used in this thesis, the combination of survival models and time series models occurs within the framework of GLMs and their extension to dynamic GLMs. I provide a detailed overview of GLMs, dynamic GLMs, and how to represent survival models as GLMs in the published GLM manuscript, which is reproduced as Appendix 3. Below, a brief overview of the key concepts is provided.

The theory of GLMs provides a unified approach for the analysis of a number of different data structures [145, 146], of which the Poisson and Binomial distributions are of particular interest to this thesis. It has long been noted that survival data may be modelled using either the Poisson or Binomial distribution [38, 129, 147, 148, 149], hence survival data may be analysed using generalised linear models. As an example, the GLM manuscript shows that the Weibull, Gompertz, log-logistic and lognormal survival models may all be written as linear GLMs. The framework of GLMs was introduced in 1972 [145], thirteen years later, this GLM framework was extended to also include time series models, with the resulting framework known as dynamic GLMs [150]. The general specification for dynamic GLMs is as follows:

$$\text{Observation model: } y_{t_i} \sim \text{distribution}(\tau_{t_i}, \mu_{t_i}]) \quad E[y_{t_i}] = \mu_{t_i}\tau_{t_i} \tag{3.1a}$$

$$\text{Response function: } \mu_{t_i} = \eta(\boldsymbol{x}_{t_i}^T \boldsymbol{\beta}_{t_i}) \tag{3.1b}$$

$$\text{Transition model: } \boldsymbol{\beta}_{t_i} = F\boldsymbol{\beta}_{t_{i-1}} + \boldsymbol{\zeta}_{t_i}, \quad \boldsymbol{\zeta}_{t_i} \sim MVN(0, \boldsymbol{Z}_{t_i}) \tag{3.1c}$$

$$\text{Initial conditions: } \boldsymbol{\beta}_0 \sim MVN(\boldsymbol{b}_0, \boldsymbol{Z}_0) \tag{3.1d}$$

With the following components:

- y_{t_i} is the outcome of interest (number of deaths in the interval starting at time t_i and end-ing at time t_{i+1}). It is assumed that there are T observations at times $t_i, i \in \{1, \ldots, T\}$. Extrapolations are required for H time units (e.g. if t_i is measured in years and extrapolations are required for 5 years, then $H = 5$) and are denoted as $\hat{y}_{T+h_i}, h_i, i \in \{1, \ldots, H\}$

- The distribution of y_{t_i} is assumed to be a member of the exponential family. For survival data either the Poisson (when modelling the hazard rate) or Binomial (when modelling the probability of death) distributions are the most relevant.

- τ_{t_i} contains information on the 'at risk' sample size in the interval $[t_i, t_{i+1})$. Let n_{t_i} be the number of people alive at time t_i and c_{t_i} be the number of censored events in the interval $[t_i, t_{i+1})$. The calculation of τ_{t_i} depends on assumptions about when events and censorings occur; this thesis uses the standard assumption that censorings occur half-way through the interval, whilst events occur at the end (see for example Chapter 2 of Tutz and Schmid [127]). Then: $\tau_{t_i} = (t_{i+1} - t_i)(n_{t_i} - c_{t_i}/2)$. Hence τ_{t_i} contains information on the at-risk sample size, the spacing between observations, and censoring (the three characteristics of survival data that cannot be accommodated by standard time series models). For a Poisson model, τ_{t_i} is included as an offset term. for a Binomial model it is included as the sample size. Use of a Binomial model requires that τ_{t_i} be an integer. For this requirement to be met, it must be assumed that censoring occurs at the end of the interval and intervals are either of the same length (so this interval length may be defined as $= 1$), or the interval lengths must all be integer multiples of the shortest length. Intervals with an unequal, non-integer, length are used in this thesis to avoid a loss of information, so the Binomial distribution is not considered further.

- $\boldsymbol{\beta}$ is a vector of parameter coefficients to be estimated from the data. In the context of a DSM, the $\boldsymbol{\beta}$ are not directly observed, so are also referred to as latent variables.

- $\boldsymbol{x}_t$ is a covariate vector, assumed known (with transpose $\boldsymbol{x}_t^T$). In general, covariates shall not be considered in this thesis as the primary focus is on modelling the evolution of the hazard function, not the impact of covariates on this function. Chapter 6 demonstrates the inclusion of a covariate to describe a time-varying treatment effect.

- $\eta()$ is a one-to-one response function which maps the linear predictor $(\boldsymbol{x}_t^T \boldsymbol{\beta}_t)$ to $f(t_i)$. Its inverse is known as the link function.

- F is a function (or transition matrix) describing how the coefficients evolve over time.
- MVN denotes a multivariate Normal distribution.
- The error term $\boldsymbol{\zeta}_{t_i}$ is assumed to be an independent and identically distributed series and $\boldsymbol{Z}_{t_i}$ is a variance-covariance matrix. In the time series literature this is also referred to as the *innovation* term; this term shall also be used in this thesis to emphasise the link between survival and time series methods.
- The initial conditions specify starting values for the latent variables (at time $t_i = 0$).

Of the terms in Equation (3.1), the observation model and response function define a standard GLM. The transition model represents a time series model; its effect is to allow the parameters of the GLM to vary over time, giving a dynamic GLM. The initial conditions define the initial values (at time zero) of the parameters.

When applying a dynamic GLM to model the hazard or survival function, resulting models are known as DSMs. For this thesis the following DSM specification is used:

$$\begin{aligned} Y_{t_i} &\sim \text{Poisson}(\exp(\beta_{1,t_i})\tau_{t_i}) \\ \beta_{1,t_i} &= \beta_{1,t_{i-1}} + \beta_{2,t_{i-1}}\phi\omega_{t_i} \\ \beta_{2,t_i} &= \beta_{2,t_{i-1}}\phi + \zeta_{t_i} \\ \zeta_{t_i} &\sim N(0, Z). \end{aligned} \tag{3.2}$$

Where $\omega_{t_i} = t_{i+1} - t_i$ is the interval width. Global models occur when the innovations are all $= 0$. To contrast with dynamic models, global models are also referred to as static models. For global models, parameter estimates do not vary over time, the only variation is in the Y_{t_i}, as described by its distribution (here Poisson). In Equation (3.2) the first line includes both the observation model and response function, the subsequent two lines are the transition model, and the last line represents the initial conditions.

For the model of Equation (3.2) β_{1,t_i} may be interpreted as the log-hazard (or average value) at time t_i. The change in the log-hazard for a unit time during the interval starting at time t_i is given by β_{2,t_i}. For extrapolations $\beta_{2,T}$ may be interpreted as the trend of a linear model. The variable $0 \leq \phi \leq 1$ is a *dampening* parameter; the lower the value of ϕ the more the impact of the extrapolated trend is reduced (dampened). Damped trend models are used in the time series literature [151], but have not been considered for DSMs before, so their use here represents a novel contribution. A particular problem arises when using damped trend models to extrapolate data with unequal time intervals as it is unclear how to define the time intervals for extrapolated periods; the approach developed for this thesis is described in Section 3.3.3. If $\phi = 1$ there is no dampening, resulting models are referred to as *local trend* models in this thesis.

In theory both β_{1,t_i} and β_{2,t_i} may vary over time. In Equation (3.2) only the β_{2,t_i} vary. This is for the following three reasons:

1. Do to their flexibility, within-sample estimates of the hazard function are in general very similar for DSMs with and without variation in the β_{1,t_i}.

2. Based on experience with DSMs (obtained during this thesis), allowing both the level and trend to vary led to increased variability of hazard estimates. This increased variation was particularly pronounced at the end of follow-up due to the small sample remaining at-risk.

3. When extrapolating survival data, it is typically expected that there will be a trend in the extrapolations (for example, due to the effects of ageing or frailty). Hence, it was decided that it was important to base extrapolations on the most recent estimates of trend. As such, if only one of $(\beta_{1,t_i}, \beta_{2,t_i})$ is allowed to vary, it was deemed important to let the β_{2,t_i} vary.

The modelling of both level and trend as time series is considered in the simulation study of Chapter 4.

Three models of particular interest arise from Equation (3.2):

Local level model: The product $\beta_{2,t_{i-1}}\phi$ is set to zero so there is only one latent variable: β_{1,t_i}. This is modelled as a time series following a random-walk, so the transition equation becomes $\beta_{1,t_i} = \beta_{1,t_{i-1}} + \zeta_{t_i}$. Extrapolations are a constant value $\hat{y}_{T+h_i} = \exp(\beta_{1,T})$

Local trend model: This sets $\phi = 1$. The trend is modelled as a random-walk. Extrapolations follow a linear trend in the log-hazard: $\hat{y}_{T+h_i} = \exp(\beta_{1,T} + h_i\beta_{2,T})$.

Damped trend model: This estimates all of the parameters in Equation (3.2). The trend follows a first-order autoregressive process. Extrapolations follow a linear trend in the short term but are gradually damped so that in the long-term extrapolations are constant: $\hat{y}_{T+h_i} = \exp(\beta_{1,T} + \sum_{i=1}^{h_i}(\phi_2^i)\beta_{2,T})$. This model may also be viewed as a weighted average of a local level and a local trend model, with the weighting given by ϕ [151].

For a derivation of the extrapolations from each model, see Appendix 1 ("Model specification for dynamic survival models"). As detailed above, the use of DSMs leads to explicit analytical formula for generating extrapolations that have intuitive interpretations relating to the anticipated future behaviour of the hazard function. This is in contrast to the majority of the alternative survival models for which no such intuitive interpretation is possible. This represents an important advantage of DSMs when used for extrapolation as it may be possible to obtain clinical input into the likely qualitative long-term behaviour of the hazard function and encode this knowledge via the appropriate choice of DSM, with alternative assumptions about the long-term hazard behaviour tested in sensitivity analyses which use other DSMs.

The local trend model is a special case of the damped trend model. When modelling a trend, one approach is to work with only the damped trend model, as the local trend will be recovered if it is appropriate. This approach is not pursued, as a key theme in this thesis is the importance of clinical input in model choice: the damped and local trend models make different assumptions about the long-term behaviour of the hazard function; it is argued that the choice between these may be based on subject-matter considerations. For implementation, when $\phi = 0$ it is not possible to identify a unique value for β_{2,t_i} (any value may be used, as it is multiplied by zero). This makes estimation of the damped trend model problematic; in practice it is recommended that $0.8 \leq \phi \leq 1$ (see Chapter 7 of Hyndman and Athanasopoulos [152]). For this thesis the less restrictive range $0.7 \leq \phi \leq 1$ is used as use of this lower range allowed for more flexibility whilst not affecting model convergence.

3.2 Current state of the art: a pearl-growing review

The aim of this review was to identify papers that described methodological aspects of DSMs in order to identify the latest methodological developments. A broad and comprehensive review of the methodological literature was conducted in the previous Chapter to identify methods for the extrapolation of survival data. This comprehensive review did not identify any studies that used DSMs, which emphasises the importance of conducting an additional review specific to DSMs. The review of this chapter was developed as a focused search of the methodological literature. To ensure that the review was focused on DSMs, a pearl-growing strategy was adopted [51]. Pearl-growing methods use an initial search to identify a set of 'pearl' articles. These pearls are then used to expand the initial results by both checking the reference lists of the pearls and searching for articles that have cited the pearls. Pearls may also include authors; results are then expanded by searching for additional publications by the authors. Searches may be iterative if the expanded results include new pearls. The application of pearl-growing methods in this review is described in the following section. The decision as to which articles were deemed to be pearls was based on an informal assessment of the quality of the article (such as if sufficient details were provided on the methods employed, and if it provided advances beyond previously published literature), and how relevant the subject of the article was to this thesis. For example, it was decided that for this thesis parameter estimation would be based on Hamiltonian Monte Carlo (see Section 3.3.2 for details and justification). As such, studies which focused on alternative methods for parameter estimation were retained but not marked as pearls.

3.2.1 Methods

The pearl-growing review had the following components:

1. Key-word searches. The phrase "Dynamic survival model" was searched for in both WoS and Google Scholar.
2. From (1), identify an initial set of pearl articles and pearl authors.
3. Perform citation searches of pearl articles in WoS.
4. Identify publications for pearl authors in WoS.

5. Search the reference lists of pearls for additional articles.
6. From (3) to (5), identify any additional pearls (articles or authors).
7. Repeat steps (3) to (6) until no new pearls are identified.

The first component (initial search) included searching two databases. The intention of this was to broaden the identified articles. Subsequent searching (components 3 and 4) were intended to be more focused, so only one database was considered. All searches were performed on the 28th of February 2018.

This search had the following inclusion and exclusion criteria:

- Papers that described a methodological aspect of a DSM were included.
- Papers that only described an application of DSM were excluded.
- There was no date restriction.
- Papers not in English were excluded.
- Grey literature was included if it provided sufficient methodological details.

3.2.2 Results

The two databases (WoS and Google Scholar) were searched separately. Fifteen articles were returned by the WoS search, of which six were retained as relevant studies [40, 144, 153, 154, 155, 156]. The search of Google Scholar returned 83 unique studies, of which seven were retained. These seven included the six identified by WoS, along with one additional study [39]. Of these seven articles, three were judged to be pearl articles, and the first authors of these articles were identified as pearl authors [40, 144, 153]. Four articles were not deemed to be pearl articles, as they focused on enhancements to DSMs that were not of direct relevance to this study. These enhancements were:

- Incorporating spatial effects [154].
- Use of auxiliary mixture sampling for parameter estimation [155].

- Incorporating the mechanics of kinematics within parameter estimation [156].
- Using a random time grid to measure the interval width [39].

Round one of pearl-growing led to the identification of a further five relevant articles: one via citation searching [157], and four from checking reference lists [128, 143, 158, 159]. Of these five articles, three were judged to be pearls, with their first authors identified as pearl authors (two studies had the same first author, so there were two pearl authors). One article was deemed to not be a pearl as it focused on methods for parameter estimation [157], whilst the other article did not describe any enhancements to the DSM beyond those described in the studies initially identified [159]. Of note, for one pearl author identified during the initial review (Jianghua He [144]), searching for author publications provided 2,745 results. After sifting the first 100 results, the majority appeared to be from different authors (with the same name). Instead, the author's personal website was used to identify publications: `https://goo.gl/DW2P6S`.

During round two of pearl-growing, citation searching led to seven relevant articles [160, 161, 162, 163, 164, 165, 166]. A further five articles were identified via author publications [167, 168, 169, 170, 171], and one additional article was found via reference list searching [172]. In total 13 relevant articles were identified. None of these 13 were deemed to be pearls. For eight articles, there were no enhancements to the DSM beyond those described by articles previously identified [160, 161, 164, 165, 168, 169, 170, 172].

Three studies described enhancements that were beyond the scope of this thesis. They were:

- To replace the transition model with a hidden Markov chain [162].
- The analysis of multivariate outcomes [163].
- The use of distributions other than the Normal in the transition model of Equation (3.1c) [167].

In addition, two studies described the use of DSMs when a proportion of the sample was assumed to be ‘cured‘ and so could not experience the outcome of interest [166, 171]. However, these studies both assumed that cured individuals never died, so their usefulness to this thesis was limited.

As no further pearls were identified, the pearl-growing review ceased after three rounds of searching (the initial search, and two rounds of pearl-growing). In summary, the initial searches identified seven relevant articles, whilst pearl-growing led to identifying an additional 18 relevant articles. The sifting process is illustrated in Figure 3.1 (initial search and round 1 of pearl-growing) and Figure 3.2 (round 2 of pearl-growing). Where the reason for exclusion is 'application', the corresponding article may or may not have used a DSM (this was not checked). The following section provides an overview of the methods employed in the 25 relevant articles.

Figure 3.1: PRISMA flow diagram for the DSM: initial review and round 1 of pearl growing

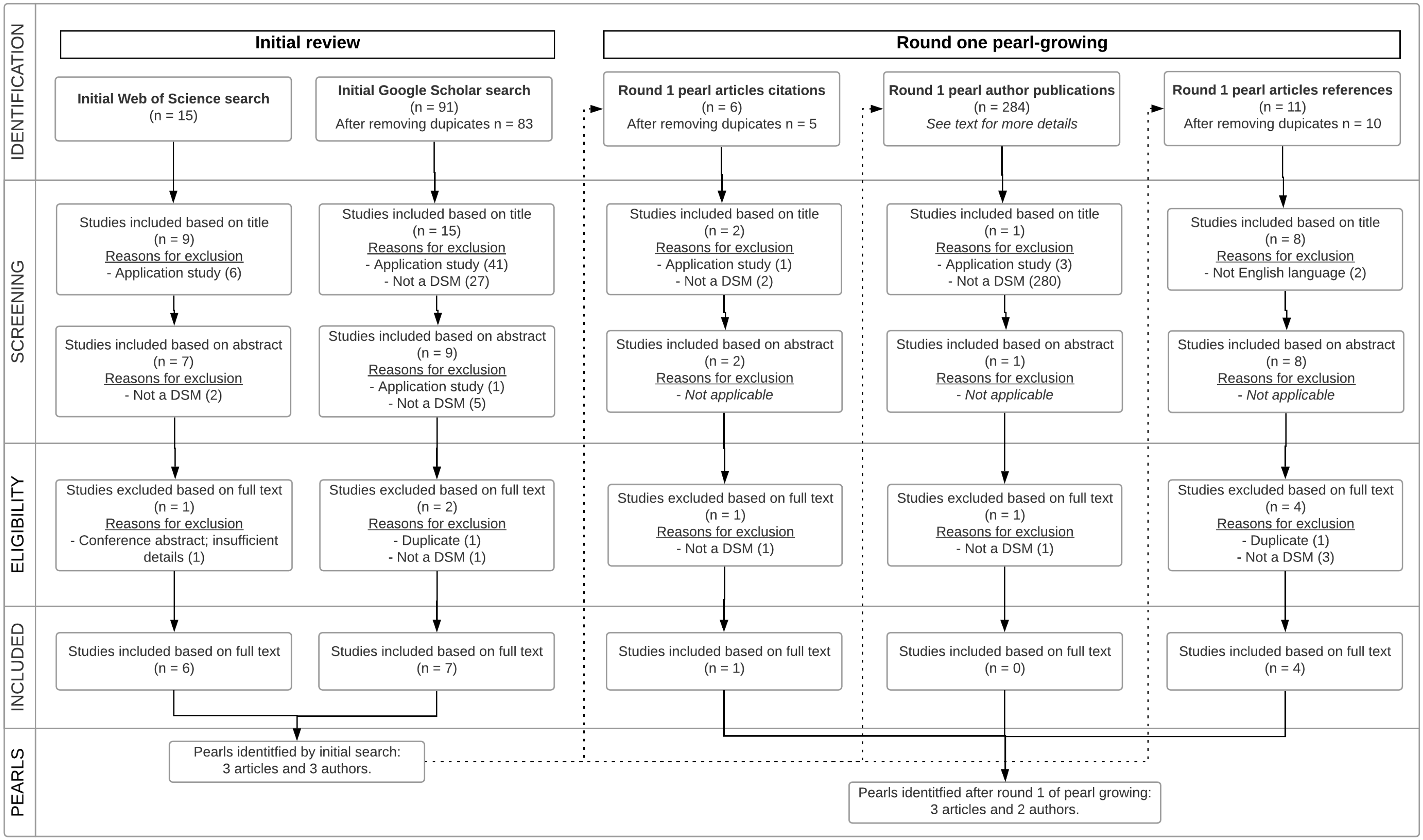
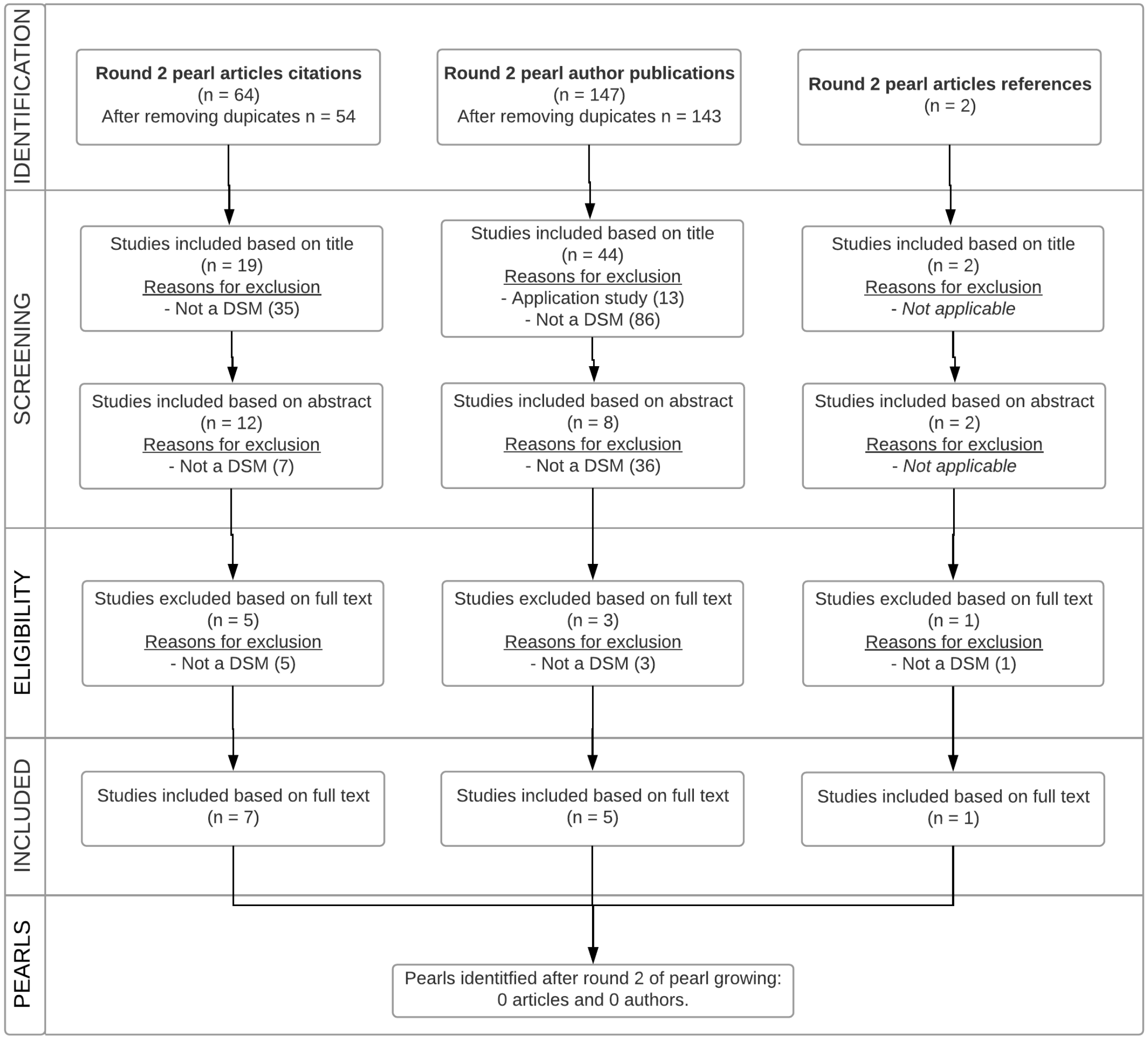

IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
PEARLS
Initial review
Round one pearl-growing
Initial Web of Science search (n = 15)
Initial Google Scholar search (n = 91) After removing dupicates n = 83
Round 1 pearl articles citations (n = 6) After removing dupicates n = 5
Round 1 pearl author publications (n = 284) See text for more details
Round 1 pearl articles references (n = 11) After removing dupicates n = 10
Studies included based on title (n = 9) Reasons for exclusion - Application study (6)
Studies included based on title (n = 15) Reasons for exclusion - Application study (41) - Not a DSM (27)
Studies included based on title (n = 2) Reasons for exclusion - Application study (1) - Not a DSM (2)
Studies included based on title (n = 1) Reasons for exclusion - Application study (3) - Not a DSM (280)
Studies included based on title (n = 8) Reasons for exclusion - Not English language (2)
Studies included based on abstract (n = 7) Reasons for exclusion - Not a DSM (2)
Studies included based on abstract (n = 9) Reasons for exclusion - Application study (1) - Not a DSM (5)
Studies included based on abstract (n = 2) Reasons for exclusion - Not applicable
Studies included based on abstract (n = 1) Reasons for exclusion - Not applicable
Studies included based on abstract (n = 8) Reasons for exclusion - Not applicable
Studies excluded based on full text (n = 1) Reasons for exclusion - Conference abstract; insufficient details (1)
Studies excluded based on full text (n = 2) Reasons for exclusion - Duplicate (1) - Not a DSM (1)
Studies excluded based on full text (n = 1) Reasons for exclusion - Not a DSM (1)
Studies excluded based on full text (n = 1) Reasons for exclusion - Not a DSM (1)
Studies excluded based on full text (n = 4) Reasons for exclusion - Duplicate (1) - Not a DSM (3)
Studies included based on full text (n = 6)
Studies included based on full text (n = 7)
Studies included based on full text (n = 1)
Studies included based on full text (n = 0)
Studies included based on full text (n = 4)
Pearls identitfied by initial search: 3 articles and 3 authors.
Pearls identitfied after round 1 of pearl growing: 3 articles and 2 authors.

Figure 3.2: PRISMA flow diagram for the DSM: round 2 of pearl growing


IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
PEARLS
Round 2 pearl articles citations (n = 64) After removing dupicates n = 54
Round 2 pearl author publications (n = 147) After removing dupicates n = 143
Round 2 pearl articles references (n = 2)
Studies included based on title (n = 19) Reasons for exclusion - Not a DSM (35)
Studies included based on title (n = 44) Reasons for exclusion - Application study (13) - Not a DSM (86)
Studies included based on title (n = 2) Reasons for exclusion - *Not applicable*
Studies included based on abstract (n = 12) Reasons for exclusion - Not a DSM (7)
Studies included based on abstract (n = 8) Reasons for exclusion - Not a DSM (36)
Studies included based on abstract (n = 2) Reasons for exclusion - *Not applicable*
Studies excluded based on full text (n = 5) Reasons for exclusion - Not a DSM (5)
Studies excluded based on full text (n = 3) Reasons for exclusion - Not a DSM (3)
Studies excluded based on full text (n = 1) Reasons for exclusion - Not a DSM (1)
Studies included based on full text (n = 7)
Studies included based on full text (n = 5)
Studies included based on full text (n = 1)
Pearls identitfied after round 2 of pearl growing: 0 articles and 0 authors.

Methodological details

Estimation methods

All 25 of the identified articles used a Bayesian framework for parameter estimation. The methods described for estimating the parameters of DSMs have evolved over-time. Initial approaches used *linear Bayes* [128] and *posterior mode estimation* [143]. The use of *Markov chain Monte Carlo (MCMC) algorithms* is now the most popular approach, being used in 71% of the identified articles that were published since 2000 (MCMC methods for DSMs were first described in 1997 [168]). Of note, the two authors who introduced linear Bayes and posterior mode estimation both used MCMC for later publications [154, 168, 170, 171]. These three estimation methods are briefly described below; further details are provided in Appendix 1 ("Parameter estimation for dynamic models").

Linear Bayes: Estimation with linear Bayes makes no distributional assumptions about the model parameters. Instead, two statistical moments of the distribution are estimated: the mean and variance [144]. It is further assumed that the innovation variances ($\boldsymbol{\zeta}_{t_i}$) follow an inverse gamma distribution, which facilitates calculations via conjugate analyses. This allows for analytical expressions to sequentially update estimates of the mean and variance over time. Full details are described by a number of authors [39, 128, 144]. The linear Bayes approach was the first described method for estimating a DSM [128]. Compared to the other methods it is relatively easy to implement and quick to run [144].

There are two main limitations with the linear Bayes approach. First, it requires that the innovation variances are known but provides no formal method for estimating these [143, 144]. Secondly, the mean and variance may not provide an adequate description of the full distribution of model parameters if they are non-linear [173, 174]. Linear Bayes estimation was the first method to be described for DSMs in 1987 [175]. Due to its limitations, as other methods for parameter estimation became available (notably MCMC in 1997 [168]), it declined in popularity: five publications were in the years 1991 to 1997, with the next publication in 2010. This 2010 publication compared linear Bayes with MCMC and concluded that use of MCMC provided results that were similar to the use of linear Bayes with optimal (known) innovation variances [144]. In practice the error variances will be unknown, so linear Bayes estimates will provide worse results than MCMC.

Posterior mode estimation: This approach obtains an approximate analytical expression for the likelihood, along with an algorithm for estimating the parameters that maximise this approximate likelihood [143]. As suggested by the name, posterior mode estimation is based on estimating the mode of the posterior distribution of the model parameters: the mode is the most probable value and can by approximated by a Normal distribution. Neither starting values nor innovation variances can be estimated using this algorithm; estimates of these are obtained via a separate expectation-maximisation algorithm. A benefit of posterior mode estimation is that it may be seen as a natural extension of model estimation methods for a DSM which used a Normal distribution in place of the Poisson [143] (the Kalman filter; see Appendix 1 for more details). It has also been noted that this approach may be motivated via a non-Bayesian framework. Estimation is then based on maximising a penalised likelihood, typically using the extended Kalman filter [158].

As with the linear Bayes method, use of posterior mode estimation provides analytical expressions for sequentially updating estimates of the mean and variance of the states. An advantage over linear Bayes is that there is no need to specify the innovation variances, as these may be estimated.

Posterior mode estimation shares the same limitation as linear Bayes in that only the mean and variance of a distribution is specified. Further, the mode may be a limited summary measure for non-linear distributions [176], particularly if there are regions of sparse data (such as at the end of follow-up) [168]. Posterior mode estimation also underestimates uncertainty [168]. Another limitation is that estimation will typically be slower than for linear Bayes [143]. After the introduction of MCMC methods for DSM parameter estimation in 1997, only one further methodological paper using posterior mode estimation was identified: a publication in 2000 which described an extension to incorporate random effects [153].

Markov chain Monte Carlo: Monte Carlo simulation techniques are a broad range of methods that may be used to estimate parameters of interest and their distributions via simulation. A Markov chain is a sequence of random variables with the property that at any iteration, the current distribution of the random variable will only depend on its previous value [177]. During each iteration, the random variable is updated according to an algorithm so that eventually the Markov chain will converge to the target distribution, which for a DSM is the distribution of parameter estimates. As such, simulations from the target distribution may be used to describe the posterior distribution of the DSM parameters. The majority of applications of MCMC for a DSM used a combination of Gibbs sampling and Metropolis-Hastings (M-H), which are both MCMC algorithms [40, 144, 154, 166, 168]. Within the Gibbs sampler it is possible to use conjugate analyses to aid implementation by assuming that the innovation variances follow an inverse gamma distribution [178]. An alternative approach is to assume that they follow an inverse Wishart distribution [168, 179] All three methods considered so far will only provide approximate solutions. However, an advantage of MCMC over both linear Bayes and posterior mode estimation is that the number of simulations may be made arbitrary large to improve accuracy. As previously noted, a simulation study suggested that MCMC methods provide results that are as good as linear Bayes with known innovation error variances [144]. Using MCMC, a number of extensions to the standard DSM have been described. These include incorporating random effects (frailty terms) [40], spatial terms [154, 170], cure models [166, 171], and multivariate outcomes [157].

Disadvantages of MCMC methods include being more computationally intensive than the previous two approaches [156], and potential issues in converging [40, 168].

Modelling prior distributions

The key difference between dynamic and global (or static) models is the inclusion of innovation terms for the evolution of the parameters over time. Hence it is anticipated that appropriate modelling of the innovation terms will be of particular importance for DSMs. In particular, the specification of prior distribution for the innovation variance(s) may be important. Two classes of prior distribution are of interest for this thesis. Uninformative (vague) priors do not contain any information about likely values for the parameter of interest. As such, posterior distributions should not be influenced by the choice of uninformative prior distribution. Weakly informative priors convey some information about the anticipated distribution of the parameter of interest, without being overly restrictive about the range of possible values. The primary motivation behind weakly informative priors is that their additional information (compared with uninformative priors) helps model fitting algorithms to converge, whilst results should remain robust (insensitive) to the choice of weakly informative prior. A full discussion of uninformative and weakly informative prior distributions may be found in Chapter 2 of Gelman *et al.* [177].

In general, the reporting of these priors was incomplete and inconsistent in the identified literature. Only seven applications specified a prior distribution for the innovation variances. All of them used an uninformative prior, which was either an inverse gamma (I.G.) distribution for the variance (four studies [144, 155, 168, 170]) or a gamma distribution for the reciprocal of the variance, which is also known as the precision (three studies [157, 161, 165]). The following values were used: I.G.(0.001, 0.001) in two studies, I.G.(0.01, 0.01), I.G.(1, 0.005), Gamma(1, 0.01), Gamma(0.01, 0.01), Gamma(0.001, 0.001). These seven priors are all designed to be uninformative: the I.G. distributions do not have a defined variance, whilst for the gamma distributions the variances ranged from 100 (mean = 1) to 10,000 (mean = 100). Of note, whilst the gamma distribution has a defined mean and variance, it allows for zero values. As such, when used to model the precision, the implied prior for the variance does not have a mean or variance. The identified papers were published between 1997 and 2011. It is likely that the choice of prior was based on it being a conjugate prior distribution (which improves the efficiency of MCMC [180]); the I.G. is a conjugate for the variance and the gamma is a conjugate for the precision.

For the latent states the most common approach was to use an uninformative prior represented by a Normal distribution with a zero mean and large variance, such as 100 [39, 40, 155, 162] or 1000 [128, 165].

Extensions to the standard DSM

The following extensions or enhancements to a standard DSM were identified:

- Incorporating random effect (frailty) terms to model unobserved heterogeneity [40, 153].
- Incorporating spatial terms [154, 170].
- Extending to model a cured fraction (although it was assumed that the cured fraction

would never experience the outcome event) [166, 171].

- Modelling multivariate outcomes [163, 157].
- Incorporating multiple (competing) outcomes [158, 168], including correlated outcomes [181]
- Using random time grids to define the time intervals (instead of defining them in advance) [39, 40].
- Incorporating probability kinematics to improve estimation with linear Bayes [156].

Comparisons of static and dynamic models

A formal comparison of static and dynamic models was performed in four studies. Two of these assessed performance using case-studies, comparing DSMs with their equivalent static model. They both found that the use of dynamic models led to improved predictions [172, 182]. Another study used both a case-study and a simulation study to compare a DSM against a spline-based model and concluded that the spline-based model was superior [170]. The final study used a simulation study to compare a DSM against a Cox model which incorporated splines [144], and the DSM was found to out-perform the Cox-spline model. None of these comparisons used survival data. As such, there is a very limited and inconclusive body of evidence about the performance of dynamic models compared to alternative models, and it is unclear how relevant the conclusions are to the setting of HTA.

3.3 Novel methodological work on DSM for this thesis

These novel contributions fall into four sections: incorporating external data, parameter estimation with Hamiltonian Monte Carlo, use of a damped trend model for extrapolating unevenly spaced data, and assessing the sensitivity of results to the choice of prior distribution for the innovation variance. These are discussed in turn.

3.3.1 Incorporating external data

The case-study of Chapter 6 and the simulation study of Chapter 5 both incorporate external data in the form of general population life tables. The motivation for the case-study was to restrict extrapolations of the hazard so that they never fell below the age-matched general population estimate of the hazard. Incorporating this constraint provides *dynamic relative survival models* (DRSMs), as it combines DSMs with relative survival models (which were identified via the comprehensive review of Chapter 2). The methodological details of these DRSMs have not previously been described in the articles identified in the pearl-growing review of Section 3.2. The simulation study of Chapter 5 was designed to reflect treatments where a proportion of the treated population can be viewed as 'cured' from their disease and so experience future survival that is the same as the general population. These dynamic cure-fraction survival models (DCFMs) have been previously described in the DSM literature [166, 171]. However, both previous examples assumed that cured individuals would never experience the event of interest. This is implausible in the context of modelling all-cause mortality, as it implies that cured individuals would never die. Hence the extension to model a non-zero hazard for the cured fraction represents a novel extension of the DSM. These two extensions are discussed in turn and use the following terminology: $\lambda^P_{t_i}, \lambda^E_{t_i}$ denote the general population hazard and the disease-specific (excess) hazard respectively (both at time t_i).

Dynamic relative survival models

The specification of a DRSM is:

$$\begin{aligned}
Y_{t_i} &\sim \text{Poisson}\left([\exp(\beta_{1,t_i}) + \lambda^P_{t_i}]\tau_{t_i}\right) \\
\beta_{1,t_i} &= \beta_{1,t_{i-1}} + \beta_{2,t_{i-1}}\phi\omega_{t_i} \\
\beta_{2,t_i} &= \beta_{2,t_{i-1}}\phi + \zeta_{t_i} \\
\zeta_{t_i} &\sim N(0, Z).
\end{aligned} \tag{3.3}$$

When compared with the specification of a DSM in Equation (3.2) the only change is to the observation model, which now includes the term $\lambda^P_{t_i}$. Hence $\exp(\beta_{1,t_i}) = \lambda^E_{t_i}$ and so β_{1,t_i} may now be interpreted as modelling the log-excess hazard. The dynamic part of the model now relates to the evolution of the disease-specific hazard over-time. This is of particular use, as it is expected that the natural history of a disease will be driven by changes in the disease-specific hazard, which in turn will be affected by the treatment received. Choice of the specific DRSM to use may then be driven by clinical input into the likely long-term behaviour of the disease-specific hazard beyond the end of the available data. The impact of alternative assumptions may be tested by assessing alternative DRSMs in scenario analyses. There are three options considered in this thesis:

The excess hazard decreases to zero. This is modelled with a local trend model. It assumes that individuals who are still alive after a certain time-point may be viewed as 'cured' from their disease, as their subsequent survival matches that of the general population. The time until this 'cure' occurs is determined by the observed data.

There is a constant excess hazard. A constant hazard is represented by a local level model. This model is appropriate if it is assumed that individuals will always be at an increased risk of death compared to the general population, and that this risk is unlikely to vary over time.

The excess hazard decreases in the short-term, to a constant value. A damped trend model is suitable in this instance. This model assumes that individuals will never be cured from their disease but that the impact of the disease on their chances of dying decrease over time.

Note that the interpretation of the local trend and damped trend models above assumes that the extrapolated trend is decreasing. If it is increasing then these two models assume that the observed increase in the excess hazard will either persist indefinitely or increase over time to a constant value, respectively.

The DRSM of Equation (3.3) assumes that the overall observed hazard is an additive composition of the general population and disease-specific hazards. Alternative assumptions are possible [183], in particular assuming a multiplicative composition as is implemented in packages such as `flexsurv` [100]. If a multiplicative relative survival model were used (i.e. $\exp(\beta_{1,t_i})\lambda^P_{t_i}$), then for values of $\lambda^E_{t_i} < 1$, the overall hazard would be less than the general population hazard. The DRSMs used in this thesis are designed to constrain the overall hazard to never fall below the general population hazard, so only the additive specification is considered. It is also noted that multiplicative relative survival models are considered to be less biologically plausible than additive models, and also generally provide worse fits to data [184].

Dynamic cure-fraction survival models

Survival models with a cure-fraction assume that the cure occurs at the start of study follow-up. Hence the overall observed survivor and hazard functions will be weighted averages of the survivor and hazard functions for the cured and uncured sub-groups. Let $S^P_{t_i}, S^U_{t_i}$ be the survival for the cured and uncured populations, and let $\lambda^U_{t_i}$ be the hazard for the uncured population (this is not the same as the excess hazard, in contrast the hazard for the cured population is $\lambda^P_{t_i}$), all at time t_i. If the fraction of cured patients is ρ, then the observed survival and hazard functions over time are given by:

$$\begin{aligned} S_{t_i} &= \rho S^P_{t_i} + (1-\rho) S^U_{t_i} \\ \lambda_{t_i} &= \frac{\rho \lambda^P_{t_i} S^P_{t_i} + (1-\rho)\lambda^U_{t_i} S^U_{t_i}}{S_{t_i}} \end{aligned} \tag{3.4}$$

For the specification of the DCFM, $Y_{t_i} \sim \text{Poisson}(\lambda_{t_i}\tau_{t_i})$, with the response function and transition model the same as the DSM of Equation (3.2). Hence $\exp(\beta_{1,t_i}) = \lambda^U_{t_i}$, which dynamically changes over time according to the transition model. As with both the DSM and DRSM, model choice for a DCFM may be guided by clinical input, this time into the likely long-term behaviour of the hazard function for uncured patients.

For this thesis two novel extensions to DSMs to incorporate external evidence are provided. These build upon the methodology of relative survival models and cure models. Approaches for incorporating external evidence which do not use dynamic models were identified by the comprehensive reviews of Chapter 2, falling under the ‘other modelling approaches’ category. These additional approaches are beyond the scope of this thesis, but their technical details are provided in Appendix 1 for completeness as this may be a useful area for further research.

3.3.2 Estimation using Hamiltonian Monte Carlo

The current approach to parameter estimation in DSMs is via MCMC, as described in Section 3.2 and Appendix 1 (technical appendix). The construction of Markov chains typically uses a type of M-H algorithm [185]. A limitation with many M-H algorithms is their relative lack of efficiency; they do not incorporate any information about the parameter space that they are exploring. It is possible to enhance standard M-H algorithms by incorporating information about the gradient of the target density to produce more efficient algorithms known as Hamiltonian Monte Carlo (HMC) [186], which also includes the Langevin method as a special case [187]. HMC supplements the stochastic MCMC algorithm with a deterministic algorithm based on Hamiltonian dynamics. This additional process uses differential equations and may be thought of as representing the 'total energy' of the Markov chain (and may be compared to the total energy of a physical system, which is the sum of kinetic and potential energy). More detailed descriptions of these algorithms can be found else [186, 187].

The HMC algorithm is more efficient than M-H MCMC. However, it still requires 'tuning' to ensure that it provides useful results. Automated tuning is an area of ongoing research. Suggested approaches include incorporating information about the geometry of the state-space [188] and restricting the Markov chain not to visit previously explored areas of the target density [189]. This latter approach, known as the *No U-turns sampler* (NUTS) has been demonstrated to perform at least as well as HMC with manual tuning [189].

An implementation of HMC is provided in the `Stan` software; an open-source program that may be used to specify statistical models and perform Bayesian inference [190, 191]. It is a C++ program, but it can be run via other programs such as R and Matlab. It provides automatic calculation of the required differential equations (and so gradients of the target space) for a given model specification. This implementation is both more robust and more efficient than alternative (standard) MCMC methods [186, 189, 191]. `Stan` is also typically faster to run than alternative programs and can fit a wider variety of models [190]. The R interface to `Stan` is used in this thesis [192]. Hence, to my knowledge, this thesis represents the first use of both `Stan` and HMC for estimating the parameters of a DSM.

3.3.3 Generating extrapolations from a damped trend model

None of the identified DSM studies used a damped trend model. This model is used in the time series literature (for example [152] Chapter 7) but assumes equally spaced intervals. Dampening occurs at discrete time-points (once per interval), and there is no guidance on how to generate extrapolations for which future intervals are not defined. As such, the approach developed for this thesis was to apply a time series model to the observed interval widths and use this to generate extrapolated interval widths for use when generating extrapolations from a damped trend model. Formally:

1. Use a dynamic linear model (local level) to generate future values of ω_i (the interval width). This is applied to the logarithm of the ω_i (to map them onto the real line), whilst noting that the ω_i may be viewed as equally spaced. The model is then:

$$\begin{aligned} \log(\omega_{t_i}) &\sim \text{Normal}(\delta_{1,t_i}, \sigma^2) \\ \delta_{1,t_i} &= \delta_{1,t_{i-1}} + \zeta_{t_i} \\ \zeta_{t_i} &\sim N(0, Z). \end{aligned} \tag{3.5}$$

   Where the value of σ is estimated from the data.

2. Use the dynamic linear model to generate future values of $\log(\omega_{t_i}); \log(\omega_{T+1}, \omega_{T+2}, \dots)$.

3. Recursively estimate the times at which future dampening occurs as: $t_{T+i} = t_{T+i-1} + \omega_{T+i}, i > 0$.

4. Repeat the previous step until $t_{T+i} > H$

3.3.4 Choice of prior for the innovation variance

The implementation of DSMs in this thesis is Bayesian. As such, prior distributions need to be specified for each of the model parameters. When compared with standard survival models, the main added value of DSMs is that they allow parameters to vary over time. The level of this variation is controlled by the value of the innovation variance. It is hypothesised that model estimates of the hazard function may be sensitive to the correct specification of the prior distribution for the innovation variance. This hypothesis was tested in a simulation study, described in the next section.

3.4 The impact of different prior distributions for the innovation variance

Where existing studies have described the prior used, this was always either the inverse gamma, or equivalently the gamma for the inverse of the innovation variance (see the DSM literature review, Section 3.2, for more details). These priors were potentially chosen as they have a conjugacy property, which can ease computations in MCMC [180]. This thesis uses HMC in preference to MCMC, so does not need to worry about choosing conjugate prior distributions. In addition, DSMs may be viewed as hierarchical models, for which a case-study showed that use of an inverse gamma prior led to poor posterior estimates [193]. Potential alternatives that were suggested include the half-Normal and half-Cauchy distributions, as well as uniform priors with range equal to the parameter space of the prior. However, as this range is $[0, \infty]$, the non-finite upper limit means that the posterior will over-estimate the variance [193]. Two studies evaluated the impact of prior choice in a DSM. One used a simulation study to compare five different prior distributions for the standard deviation: four were inverse gamma, with parameter pairs $(0.01, 0.01)$, $(10^{-4}, 10^{-4})$, $(10^{-5}, 10^{-5})$, $(10^{-8}, 10^{-8})$, along with a uniform prior on $[0, \infty]$ [170]. The authors found that results were robust to the choice of prior. In contrast, in a separate case the authors noted that use of a Gamma$(1, 0.01)$ for the precision provided better results than a Gamma$(0.01, 0.01)$ [161].

There is no consensus on what prior distribution should be used for the innovation variance in a DSM, so a simulation study was performed with the following aims:

1. Identify rates of convergence for key model parameters.

2. Compare model estimates of the innovation variance against the true value.

3. Assess agreement between model-estimates of the log-hazard and the true values.

In addition, the impact of choice of prior distribution was also assessed using the case-studies of Section 4.1 and Chapter 6. Details on the methods for these and results are provided in Appendix 2.

3.4.1 Methods

The presentation of the methods follows published guidance [194], with separate sections for each of the key components of the simulation study.

Data-generating mechanism and estimand

Data were simulated from a local-level dynamic survival model. The starting value of the log-hazard $\beta_{t_0} = \log(0.05)$ and data were simulated for 25 evenly spaced time periods. The initial at-risk sample size was $\tau_{t_1} = 300$. For a pre-specified value of the innovation variance Z, data were simulated as follows:

1. Generate 25 innovations from $\text{Normal}(0, Z); \zeta_{t_i}, t_i \in (1, \dots, 25)$.
2. Generate 25 log-hazard values: $\beta_{t_i} = \beta_{t_i-1} + \zeta_{t_i}$.
3. Sample the number of deaths in the first time interval $y_{t_1} \sim \text{Poisson}(300 \exp(\beta_{1,t_1}))$.
4. Update the at-risk sample size: $\tau_{t_i} = \tau_{t_i-i} - y_{t_i-1}$
5. Sample the number of deaths per subsequent interval as $y_{t_i} \sim \text{Poisson}(\exp(\beta_{1,t_i})\tau_{t_i})$
6. Discard any time-intervals for which the at-risk sample = 0.

Three different values of Z were used: 0.001, 0.01, and 0.1. The focus of this study was on small values for Z, as a variance of 0 is equal to a static mode, and it was felt to be important that DSMs can identify this special case. For each value of Z, 200 datasets (n_{sim}) were simulated. The primary estimand was Z. The secondary estimand was the time-varying log-hazard β_{t_i}.

Prior distributions

Six uninformative and six informative prior distributions were considered. The informative priors were designed to be weakly informative, but the interpretation of what constitutes weak prior information is subjective, hence these priors are referred to as 'informative'. The parameterisation of all the priors is provided in Table 3.1. Uninformative priors were defined as those with an undefined variance. All of the priors are for the variance, except the first two priors which are for the standard deviation. The choice of distributions was based on those considered in the literature (either for DSMs, or for hierarchical models in general), as well as distributions which have support on the positive line, so are appropriate for modelling variances. Initial values were not specified for any of the parameters, instead the `Stan` defaults were used.

Table 3.1: Prior distributions used for the variance (apart from prior IDs 1 and 2).

ID	**Uninformative prior distributions**	**Mean (variance)**
1	Uniform$(0, \infty)$ for standard deviation	Undefined (undefined)
2	Uniform$(-\infty, \infty)$ for log(standard deviation)	Undefined (undefined)
3	Uniform$(0, \infty)$	Undefined (undefined)
4	Inv.Gamma$(0.01, 0.01)$	Undefined (undefined)
5	Inv.Gamma$(1, 0.005)$	Undefined (undefined)
6	Half-Cauchy$(0, 0.3)$	Undefined (undefined)
	Informative prior distributions	
7	Half-normal$(0, 0.3)$	0.44 (0.11)
8	Half-normal$(0, 5)$	1.78 (1.82)
9	Gamma$(0.01, 0.01)$	1 (100)
10	Gamma$(0.02, 0.1)$	0.2 (2)
11	Gamma$(2, 1)$	2 (2)
12	Lognormal$(0, 1)$	1.65 (1.72)

Performance measures

For estimates of Z, both mean squared error (MSE) and bias were calculated as [194]:

$$\text{MSE} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_{t_i} - \theta)^2 \tag{3.6}$$

$$\text{Bias} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_{t_i} - \theta) \tag{3.7}$$

where $\hat{\theta}_{t_i}$ is the model-based estimate of the estimand $\hat{\theta}_i$. For both performance measures, lower values indicate better model performance.

The performance of model estimates of the log-hazard were compared visually using Bland-Altman plots [195]. This plots the average of the model estimates and the truth on the x-axis, against the difference on the y-axis. For perfect fit, the difference should be zero. There should also be no pattern in the differences by average measurement.

A final performance measure for each model was whether it converged during parameter estimation. Four chains were used during model estimation. Convergence was assessed using the

$\hat{R}$ statistic (also referred to as the potential scale reduction, see Chapter 11 of Gelman *et al.* [177]). This is the ratio of the between-chain variance to the within-chain variance. A value of $\hat{R} \approx 1$ implies that the fitting algorithm may have converged, whilst larger values suggest that more iterations may be required. There are no specific rules to identify when $\hat{R}$ is too large, although an upper bound of 1.01 has been recently suggested [196]. As such, two upper bounds are considered for this study: 1.01 and 1.20 (as an arbitrary less strict upper bound). In addition, trace plots were used to visualise assess convergence and identify if a lack of convergence was associated with particular parameter values [180]

3.4.2 Results and discussion

Summary values of the distribution of $\hat{R}$ values for each prior distribution are provided in Table 3.2. A clear lack of convergence may be seen for the second prior distribution (uniform for the log of the standard deviation), with a mean $\hat{R}$ value much larger than for the other distributions. In addition, over 70% of simulations had an $\hat{R}$ value greater than the cut-off of 1.01, and almost half had values greater than 1.2. Hence, the second prior distribution was not considered further. Use of an I.G. prior resulted in models with good convergence; for the two prior distributions considered, mean and max $\hat{R}$ values were 1.00 and 1.05 for both. These were the lowest observed max $\hat{R}$ values for all of the distributions, along with the Gamma(2, 1) prior, but this prior had more values > 1.01. As such, the two I.G. priors resulted in the smallest distribution of $\hat{R}$ values. The first prior distribution (uniform for standard deviation) had one very large $\hat{R}$ for $Z = 0.001$. Apart from this it has similar performance to the other priors.

To summarise, with the exception of the second prior, all of the prior distributions showed acceptable mean $\hat{R}$ values, which ranged from 1.00 to 1.01. There was no notable difference between uninformative and informative priors and only some minor differences in the tail behaviour of the $\hat{R}$ statistic between the different priors.

A comparison of model estimates of the innovation variances against the known true value is provided in Figure 3.3. The corresponding MSEs for these estimates are provided in Figure 3.4 and Table 3.3. Results for the bias were very similar and are not reported. The first prior has large MSE and bias for $Z = 0.001$, due to the previously noted lack of fit, so is not shown

Table 3.2: Convergence diagnostics ($\hat{R}$) for the prior distributions.

ID	**Uninformative prior distributions**	**Mean $\hat{R}$**	**Max $\hat{R}$**	**$\hat{R} > 1.2$**	**$\hat{R} > 1.01$**
1	Uniform$(0, \infty)$ for standard deviation	1.01	3.11	0.20%	12%
2	Uniform$(-\infty, \infty)$ for log(standard deviation)	1.43	11.71	48%	71%
3	Uniform$(0, \infty)$	1.00	1.08	0%	14%
4	Inv.Gamma$(0.01, 0.01)$	1.00	1.05	0%	9%
5	Inv.Gamma$(1, 0.005)$	1.00	1.05	0%	8%
6	Half-Cauchy$(0, 0.3)$	1.00	1.07	0%	11%
ID	**Informative prior distributions**	**Mean** $\hat{R}$	**Max** $\hat{R}$	$\hat{R} > 1.2$	$\hat{R} > 1.01$
7	Half-normal$(0, 0.3)$	1.00	1.06	0%	9%
8	Half-normal$(0, 5)$	1.00	1.07	0%	13%
9	Gamma$(0.01, 0.01)$	1.01	1.24	0.20%	15%
10	Gamma$(0.02, 0.1)$	1.01	1.13	0%	16%
11	Gamma$(2, 1)$	1.00	1.05	0%	14%
12	Lognormal$(0, 1)$	1.00	1.07	0%	8%

in Figure 3.4. Overall, the inverse gamma $(1, 0.005)$ had the best performance as it had the lowest MSE values for two of the prior values considered ($Z = 0.01, 0.1$) and the third lowest MSE for the other prior ($Z = 0.001$). Two of the gamma priors (with parameters (0.02, 0.1) and (0.01, 0.01)) also had very good results, as their MSE was always amongst the three lowest values (along with the previously mentioned inverse gamma). Of the uninformative priors, use of either an inverse gamma or half-Cauchy resulted in better performance than use of a uniform distribution. The largest errors (and biases) were observed for the gamma (2, 1) and lognormal (0, 1) priors.

A visual comparison of model estimates of the log-hazard against the true values is provided via Bland-Altman plots in Figure 3.5. These all showed close agreement, except for the first prior for variance = 0.001, as previously noted. There was also, for all the prior distributions, a slight increase in the spread of differences for the largest values of the average log-hazard. This is to be expected, as larger hazard values imply that almost all the at-risk sample will die, which reduces the sample size and so will make hazard estimates less precise.

Based on MSE and bias, results from the simulation studies supported the use of either an inverse gamma prior distribution or a gamma distribution with a large variance relative to its mean. However, with regards to impact on predicted hazard values, any of the prior distributions that were considered performed well, with the exception of the second prior; Uniform$(-\infty, \infty)$ for the logarithm of the standard deviation. This robustness of log-hazard estimates to the choice of prior was also observed in the two case-studies (see Appendix 2). Together, the results of the simulation study and case-studies suggest that the specific choice of prior is primarily of importance if the resulting model estimation fails to converge. Hence, as it had the lowest MSE and bias in the simulation studies, the inverse gamma (1, 0.005) is adopted as the default prior distribution for the innovation variance in this thesis. Other prior choices would have been considered if model estimation failed to converge, although this never occurred for the analyses presented in this thesis.

Table 3.3: Mean squared error (x1,000) with 95% confidence intervals: estimates of innovation variance.

ID	**Uninformative priors**	**Z = 0.001**	**Z = 0.01**	**Z = 0.1**
1	Uniform$(0, \infty)$ std. dev.	9.54 (7.64 to 11.44)	18.45 (15.72 to 21.18)	55.4 (52.22 to 58.58)
3	Uniform$(0, \infty)$	1.21 (0.95 to 1.47)	2.43 (1.76 to 3.11)	1.71 (1.3 to 2.12)
4	Inv.Gamma$(0.01, 0.01)$	0.54 (0.46 to 0.62)	0.73 (0.52 to 0.94)	1.12 (0.86 to 1.39)
5	Inv.Gamma$(1, 0.005)$	0.05 (0.04 to 0.05)	0.06 (0.03 to 0.09)	0.82 (0.64 to 0.99)
6	Half-Cauchy$(0, 0.3)$	0.94 (0.77 to 1.1)	1.65 (1.24 to 2.05)	1.32 (1.02 to 1.62)
ID	**Informative priors**	**Z = 0.001**	**Z = 0.01**	**Z = 0.1**
7	Half-normal$(0, 0.3)$	1.01 (0.82 to 1.2)	1.74 (1.31 to 2.18)	1.4 (1.09 to 1.7)
8	Half-normal$(0, 5)$	1.17 (0.92 to 1.42)	2.34 (1.69 to 2.99)	1.66 (1.27 to 2.05)
9	Gamma$(0.01, 0.01)$	0 (0 to 0)	0.22 (0.12 to 0.32)	1.11 (0.82 to 1.39)
10	Gamma$(0.02, 0.1)$	0.01 (0 to 0.01)	0.23 (0.14 to 0.32)	1.1 (0.84 to 1.35)
11	Gamma$(2, 1)$	6.26 (5.45 to 7.07)	8.49 (7.05 to 9.93)	2.67 (2.08 to 3.26)
12	Lognormal$(0, 1)$	10.8 (9.99 to 11.6)	12.07 (10.87 to 13.27)	2.5 (2.01 to 2.99)

Figure 3.3: Comparison of model estimates of the innovation variance and the true values. Black line represents truth.

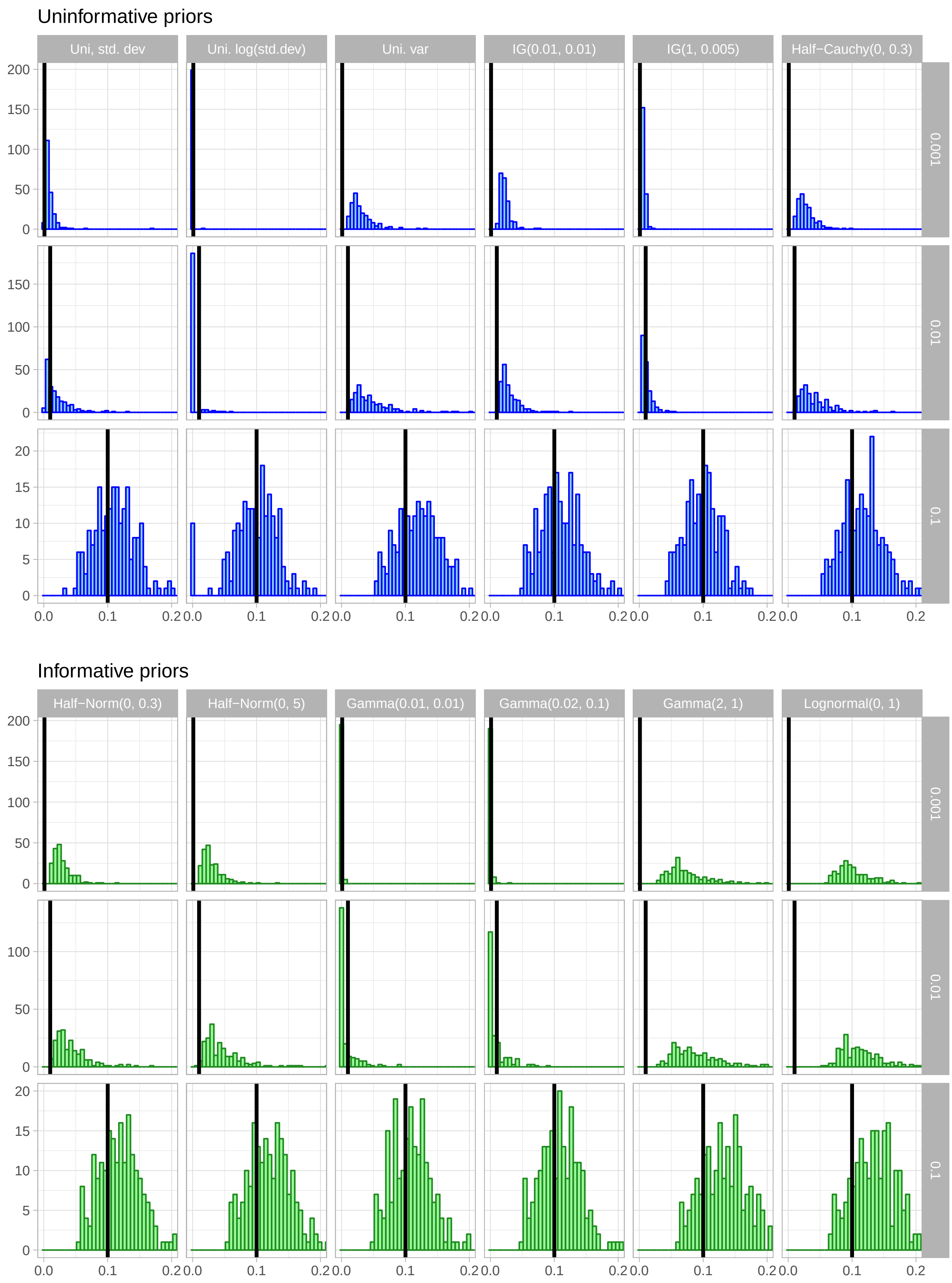


Figure 3.4: Mean-squared error (top panel) and bias (bottom panel) for the 12 prior distributions.

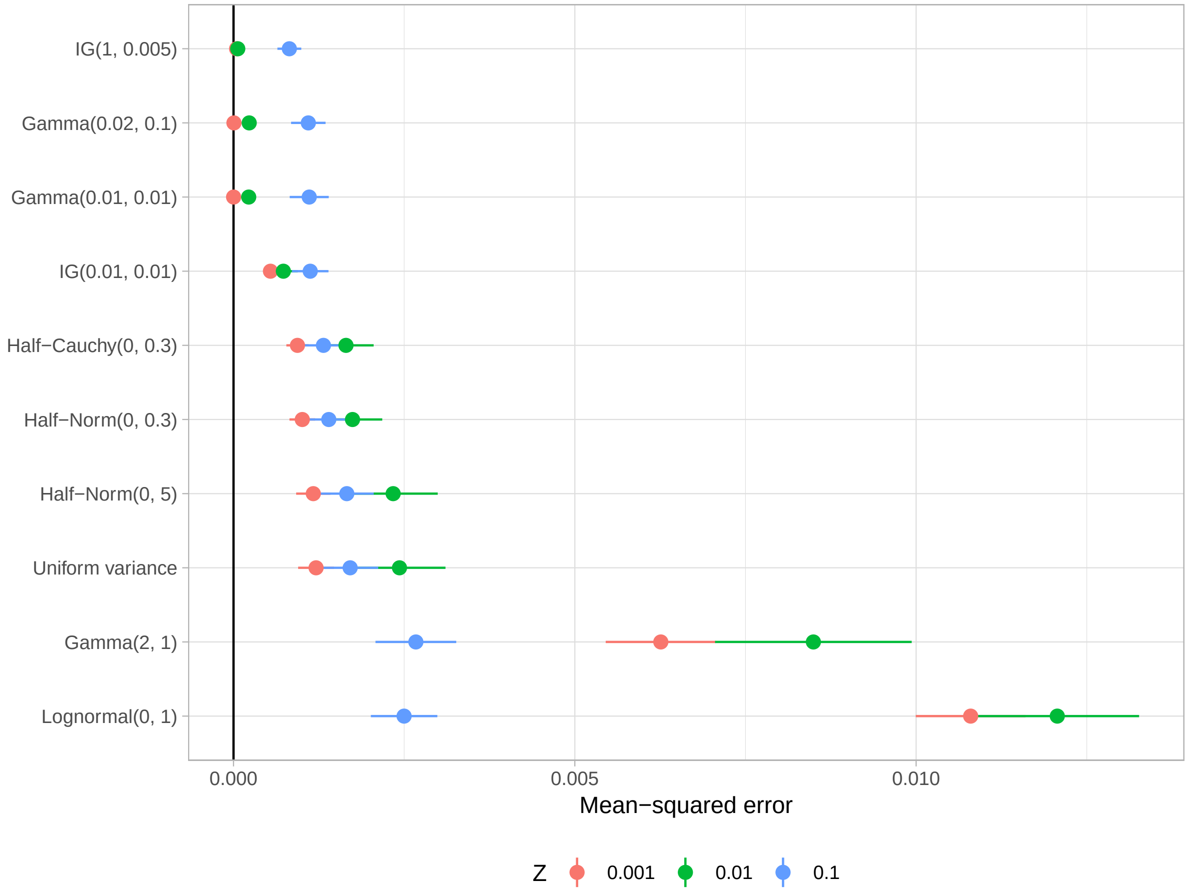
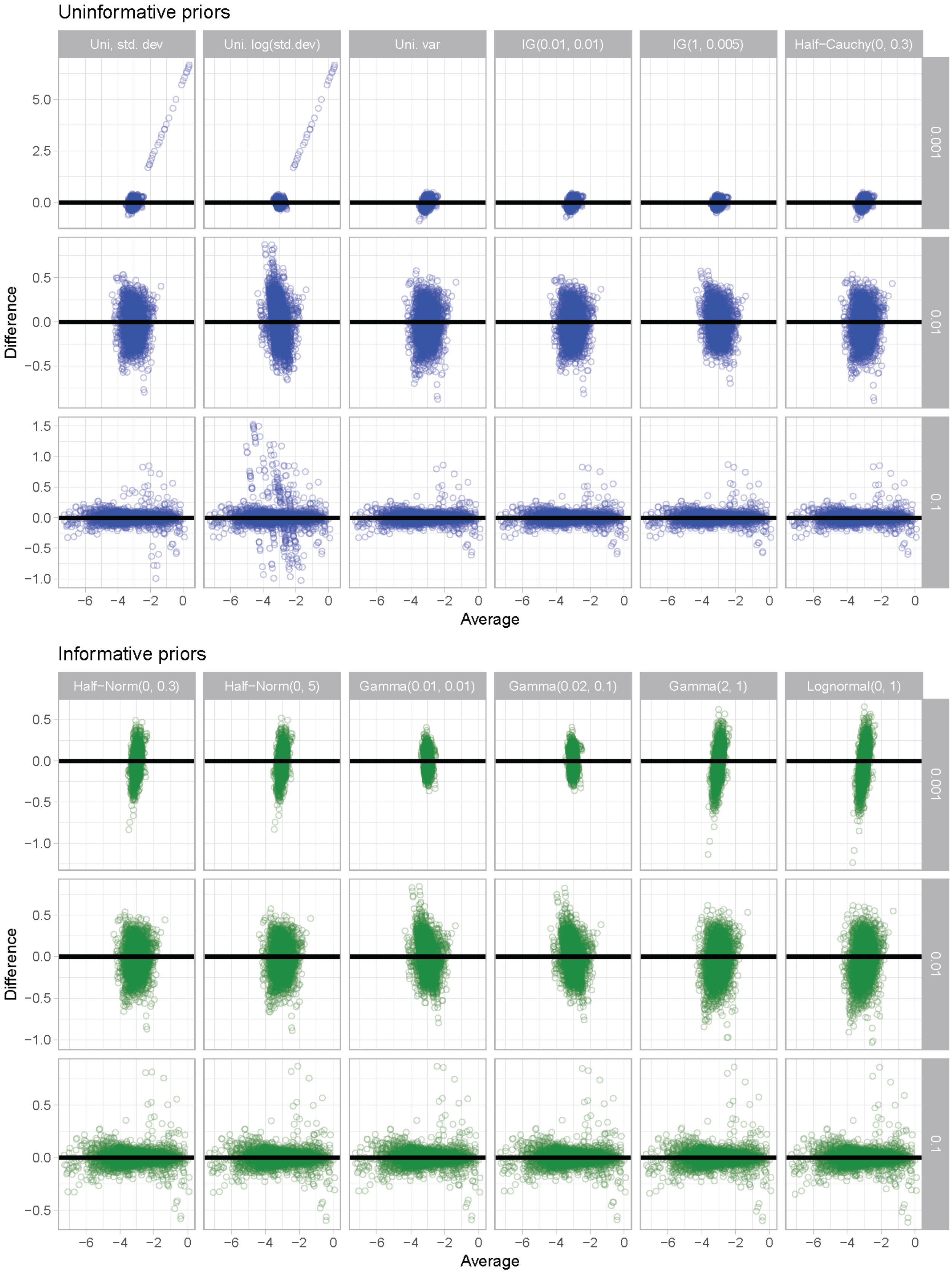


Figure 3.5: Bland-Altman plot of the log-hazard: model estimates vs truth (rows = value of Z).



3.5 Conclusions

The union of time series and survival models is known as DSMs, which are in turn a special case of dynamic GLMs. This chapter has introduced the specification of DSMs, which has two main components. The first is a model for the observed data (observation model), which is a GLM for survival data. The second is the transition model, which defines a time series model for the temporal evolution of the parameters of the observation model. It is important to ensure that the application of DSMs in this thesis builds upon the current state of the art for DSMs, which was identified via a pearl-growing literature review. This review identified that MCMC methods were the most advanced method for parameter estimation in current implementations. The review also highlighted some key evidence gaps, as there was a paucity of comparative studies, both comparing DSMs to alternative survival models and comparing the impact of different prior distributions for the innovation variance (which defines how 'dynamic' a DSM is). Further, there were no identified articles that considered incorporating external evidence with a DSM. These evidence gaps are addressed within this thesis. Two novel extensions to incorporate external evidence in the form of general population life tables were described; one for which this external evidence was used to constrain the extrapolations (so future hazards never fell below the general population hazards) and one where the life tables were used to represent the future survival of a cured fraction of individuals. The impact of prior choice for the innovation variance was evaluated in a simulation study which compared 12 different prior distributions and found that estimates of the hazard function were generally insensitive to the choice of prior distribution, of which an inverse gamma prior gave the smallest bias and MSE.

This chapter has provided an overview of DSMs, including useful extensions to incorporate external evidence and a simulation study to inform model specification. This novel work was motivated by the results of a focused literature review to identify the current state of the art for DSMs. The previous chapter identified a set of comparator methods that may be used as alternatives to DSMs for the analysis and extrapolation of survival data. The subsequent two chapters evaluate the performance of DSMs and the comparator methods, in both a case-study and simulation studies.

Chapter 4

Comparing dynamic survival models with current and emerging practice

To be useful in a technology appraisal, survival models should combine good within-sample fit with accurate extrapolation performance. Extrapolation performance relates to the ability of a statistical model to accurately predict future observations (outcomes), given the available observations. Evaluations of both within-sample fit and extrapolation performance require that future observations (which were not used for model fitting) are available. Ideally this evidence would be obtained from real datasets, such as trials with long-term follow-up. For this thesis one such dataset was identified, and it is analysed in this chapter. There are two particular strengths of this case study. First, it enabled access to patient-level data with long-term follow-up. Secondly, there is a published interim analysis which serves as a natural choice of data to use to generate extrapolations (which are compared against the full dataset). This case study is a trial of abiraterone acetate (henceforth referred to as abiraterone) in prostate cancer.

To avoid estimates of overall goodness of fit (within-sample and extrapolations) being driven by the quirks of a single dataset, evaluations should be conducted for multiple datasets. As there was a lack of additional suitable real datasets for this thesis, the required evidence was obtained via simulation. The advantage of simulation studies is that the 'truth' (here the true hazard function) is known, so can be used to evaluate performance. Two simulation studies were performed for this thesis. Both had the aim of assessing the overall goodness of fit - of

DSMs, current practice, and emerging practice - for realistic hazard patterns. The next chapter considers a hazard function indicative of a sample with a cured fraction. This chapter considers a multi-modal hazard function (one with two turning points), arising from a mixture Weibull model. This could represent a population of 'low-risk' and 'high-risk' individuals. An overview of the methods used to conduct the simulation study is provided, along with the study results and a discussion of their implications for this thesis. The two analyses within this chapter (the case study and the simulation study) are discussed fully in turn, followed by an overarching conclusion.

4.1 Abiraterone case-study

The objective of this case study is to evaluate the extrapolation performance of DSMs and alternative approaches when applied to real data. The data used for this case study come from the clinical trial COU-AA-301, which compared abiraterone to placebo treatment in people with castration-resistant prostate cancer previously treated with docetaxel-based chemotherapy (NCT00638690). The individual patient-level data were accessed via the Yale University Open Data Access Project [197]. The full dataset had almost complete follow-up and an early cut of the data has been published in the form of an interim analysis [198]. The approach taken here is to apply the statistical models to the early cut of the data, using the more complete data to evaluate the performance of the extrapolations.

The following subsection describes the available data, how the interim data were recreated from this, and provides an initial exploratory analysis. Subsequent subsections provide details on how the candidate extrapolation models were identified. The extrapolations from these models are then compared against the longer-term data. The final subsection for this case-study provides a discussion of the findings.

4.1.1 Data used and exploratory analysis

The pivotal trial COU-AA-301 was an international trial evaluating the clinical effectiveness of abiraterone. Patients were randomised 2:1 to receive either abiraterone or placebo, both study groups also received prednisone. Treatment was until disease progression and the primary outcome was overall survival. Between May 2008 and July 2009 1,195 patients were enrolled (abiraterone = 797, placebo = 398). A pre-specified interim analysis was published in 2011, based on 552 deaths (planned deaths = 534), with a median follow-up of 12.8 months [198]. A further analysis was published after 775 deaths occurred [199]. Following this analysis, patients who remained on treatment (125 [16%] and 18 [5%] for abiraterone and placebo respectively) were un-blinded and the placebo group crossed-over to receive abiraterone. An adjustment for treatment switching to abiraterone is not considered here due to the small numbers involved.

The data available for this study were made available by Janssen Research and Development, L.L.C. via the Yale University Open Data Access Project and represent more complete follow-up than either of the previous analyses, with 984 deaths (82.3% of the original study population). Data are for individual patients and were accessed via a secure online platform. For these data survival times (deaths or censorings) are measured as times from randomisation. As calendar times (for study entry) are not available, the interim dataset was replicated based on the following steps:

1. Find the follow-up time of the 552$^{\text{nd}}$ death in the dataset. Let this time $= T^*$.
2. Censor all follow-up times greater than T^*.
3. For newly censored times, set the follow-up time = time of last study visit.

One further difference from the original interim analysis is that 12 people did not have any event or censoring times, so were excluded from all analyses (hence $N = 1,195$–$12 = 1,185$).

A comparison of the available data and the published interim analysis is provided in Table 4.1. The replicated interim data is generally very similar to the published analysis, albeit with a slight under-estimate of median survival for the abiraterone group. Median follow-up in the full dataset is almost three times that of the interim analysis, with almost twice as many deaths. Of those who had not died in the full dataset, 155 (13.1% of the total sample) were still alive (and so censored at the last study visit day), with a further 38 patients withdrawn from the study and 6 lost to follow-up (and so censored at the recorded timings of the study-events).

Table 4.1: Comparison of cost-effectiveness results: original submission and replication.

	Published interim analysis	**Data used in this case study**	
		Interim replicated	**Full dataset**
Sample size: Abiraterone, Placebo	797, 398	791, 392	791, 392
Median follow-up	12.8	13.1	36.2
Median survival:			
Abiraterone	14.8	13.9	15.9
Placebo	10.9	11.1	11.2
Deaths: Total (Abiraterone, Placebo)	552 (46.2%) (333, 219)	552 (46.7%) (329, 223)	984 (83.2%) (651, 333)

The subsequent exploratory data analysis and model fitting uses the replicated interim analysis. An overview of the number of participants remaining in the study over time is provided in Table 4.2. For up to one year of follow-up the absolute number of patients remaining in the study is large, suggesting that hazard estimates should be sufficiently precise. There is a notable drop-off in numbers after 1 year which is mainly due to censoring, but even at 14 months participant numbers remain acceptable for both the abiraterone ($n = 199$) and placebo ($n = 47$) groups.

Table 4.2: Participants in the study (replicated interim dataset). Data are counts (percentage of starting cohort).

Time (months)	**Abiraterone (n = 791)**	**Placebo (n = 392)**
2	768 (97%)	379 (97%)
4	714 (90%)	340 (87%)
6	660 (83%)	303 (77%)
8	591 (75%)	263 (67%)
10	529 (67%)	214 (55%)
12	434 (55%)	164 (42%)
14	199 (25%)	47 (12%)

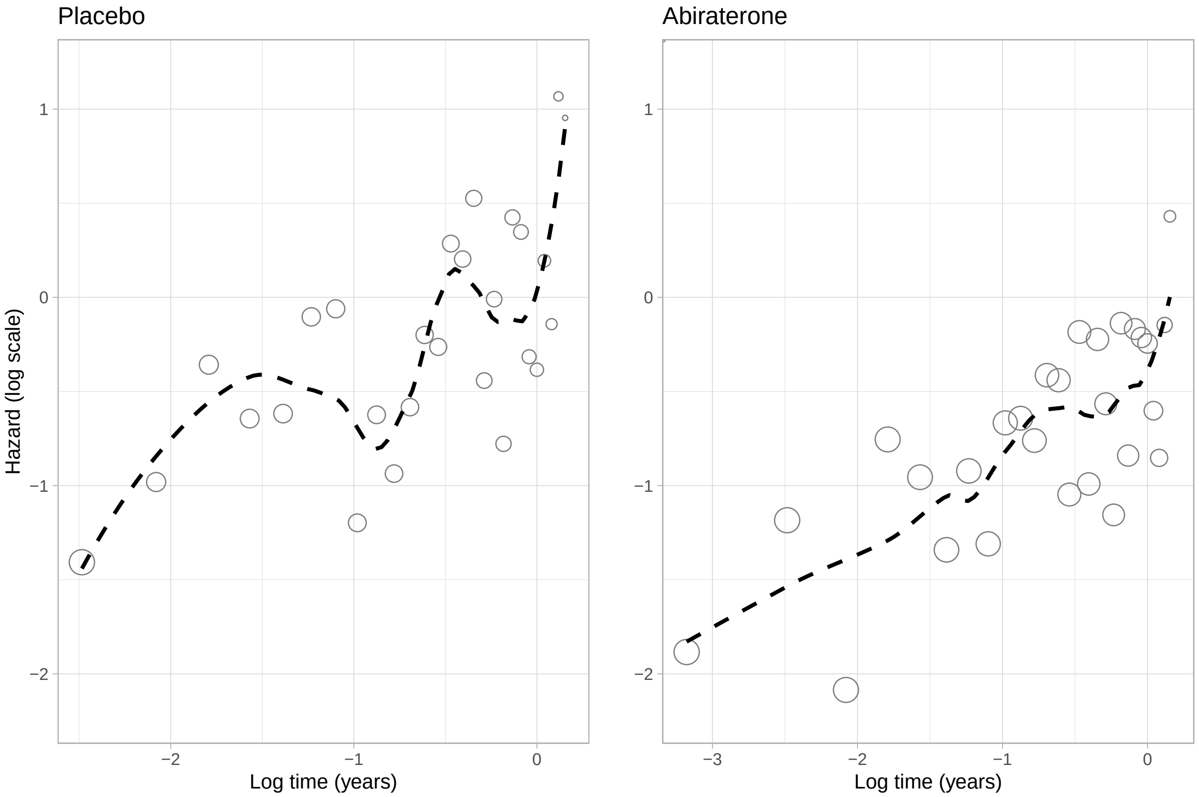
Graphs of the empirical hazard, along with a smooth non-parametric estimate, are provided in Figure 4.1 for both treatment groups. Some things to note.

- As the hazard is a rate, its value will depend on the time-intervals that are used. For this exploratory graph, fortnightly estimates of the hazard are used, to reduce the variation in the plotted hazards. For all model estimation, continuous time estimates of the hazard are used. For models using the GLM and DGLM frameworks, this uses one death per time interval, as discussed in Chapter 3.

- A loess smoother is used to obtain the smooth non-parametric estimate. This is intended to provide an informal visual aid, without making any structural assumptions. A limitation with this approach is that it does not account for the decreasing sample size over time. The loess smoother also cannot be used for extrapolation.

- Figure 4.1 plots the log-hazard against log-time. This mirrors the specification used for the majority of the models considered here. In addition, using the logarithm of time gives more visual emphasis to estimates at the start of follow-up (where there is more data). The x-axis begins slightly earlier for abiraterone than placebo as the first event occurs earlier.

Figure 4.1: Empirical estimates of the hazard function by time with a non-parametric smooth estimate

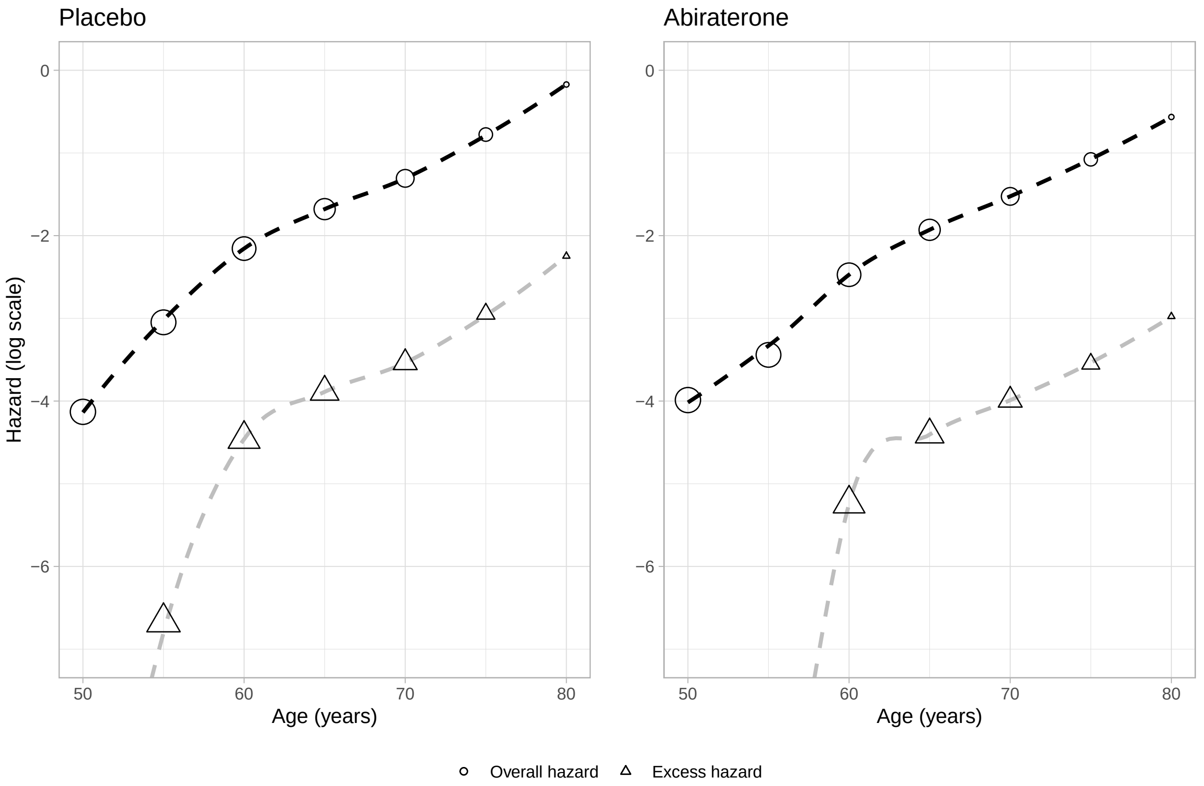


For both groups, the log-hazard follows a near-linear increase over time. This is most pronounced for the abiraterone group; for the placebo group some of the additional variation may be attributable to the smaller sample size. This suggests that, of the standard survival models, the exponential is unlikely to provide a good fit, whilst the Weibull (which is a linear model of the log-hazard against log-time) may provide a good fit. All of the other standard survival models are able to capture an increasing hazard. However, of note, neither the log-logistic nor the lognormal can model a hazard pattern that increases indefinitely. Instead, they assume that after a certain time-point the hazard function will decrease.

There are at least two potential drivers for the observed increase in the hazard of mortality over time. The first is the impact of ageing, as in general, older participants are more likely to die than younger participants. The second is disease-specific (deaths due to prostate cancer). Some insight into the impact of ageing on temporal changes in the hazard may be obtained by plotting the hazard rate against baseline age. This is not the same as estimating the temporal association between age and the hazard of death (as baseline age provides a cross-sectional measure) but it may be a useful proxy measure. In addition, differences by baseline age have been used to generate extrapolations in previous studies [23, 76] (See 'Modelling on both the age and timescales' of Appendix 2 for further details on this approach). An approximate estimate of the disease-specific (excess) hazard is also obtained by removing the age-matched general population hazard. This is obtained for the USA for the year 2008 (enrolment was May 2008 to July 2009) from the Human Mortality Database [200]. Resulting plots are in Figure 4.1.

The relationship between baseline age and hazard is very similar for both treatments. On the log-scale, there appears to be a piecewise linear relationship between age and the overall log-hazard, with a large increase for those under 60 and a smaller increase for those over 60. Before the age of 60 the excess hazard is approximately zero (hence cannot be plotted on the log-scale) and follows a linear increase after the age of 60. This suggests that ageing will lead to increases in both the overall hazard and the disease-specific hazard over time, with larger increases at the start of follow-up, due to ageing amongst those under 60. The observed changes in the overall hazard over time will depend on both the impact of ageing and disease-specific impacts.

Figure 4.2: Empirical estimates of the hazard function by baseline age with a non-parametric smooth estimate



4.1.2 Candidate extrapolation models; within-sample fit

Five classes of survival model were considered:

1. Current practice models: Eight survival models were considered: exponential, Weibull, Gompertz, gamma, log-logistic, lognormal, generalised gamma, and generalised F. Two models were chosen to be taken forwards as candidate extrapolation models.

2. RPMs: between zero and five internal knots were considered, with outcomes of both the log cumulative hazard and log cumulative odds considered, resulting in 12 potential models (when using no internal knots the two models correspond to Weibull and log-logistic models, respectively). One model was chosen for extrapolations.

3. GAMs: As model complexity is automatically penalised during model estimation, only one model was considered, which used the log of time.

4. Fractional polynomials: Both first and second-order polynomials were considered, with the use of log-time. One model was chosen for extrapolations.

5. DSMs: local trend and damped trend models were used. These have the same specification (including prior distributions) as for the simulation study of this chapter, which provides more details. Both models used a global level.

For classes one, two, and four, multiple models may be fit. The choice of model(s) to use for extrapolations was based on a combination of goodness of fit to the observed data and the plausibility of extrapolations. Goodness of fit included both visual fit and two quantitative measures: Akaike and Bayesian information criteria (AIC and BIC, respectively).

Within-sample goodness of fit values for the standard and RPMs are provided in Table 4.3. Both information criteria provide very similar values. Whilst models with lower AIC are preferred, it can be hard to evaluate meaningful differences in AIC values, as there are no guidelines for non-nested models. To alleviate this, the inverse evidence ratio (IER) is used. This is a measure of how plausible a model is, relative to the 'best' model (the model with the minimum information criteria). Let IC_m be the information criteria value (such as AIC) for model m, and IC^* be the minimum of IC_m. Then the IER for model m is defined as: $\exp(-0.5 * [IC_m - IC^*])$. The best fitting model will always have an $IC_m = 100\%$, whilst values for poorly fitting models will be close to zero [201]. IER values are calculated within each class of model. An advantage of using IERs is the relative ease of comparing models, although other methods such as linking BIC to Bayes factors may also be used [202].

Visual goodness of fit (within and out of sample) was considered for all but is only presented for the models taken forwards for extrapolation. These visualisations are provided in Figures 4.3 and 4.4 for the abiraterone and placebo groups, respectively.

The Weibull and gamma models had very similar AIC and BIC values and for both groups are the two best standard models, with IERs of at least 85% (no other standard models had values greater than 39%). Within-sample the Weibull and gamma models provide very similar estimates for both treatment arms, whilst extrapolated values are lower for the gamma. Both Weibull and gamma models were considered for extrapolation.

Table 4.3: Goodness of fit measures.

	Abiraterone				**Placebo**			
Current practice models	**AIC**	**IER**	**BIC**	**IER**	**AIC**	**IER**	**BIC**	**IER**
Weibull	1,086.3	100%	1,093.9	100%	536.3	87%	543.1	87%
Gamma	1,086.6	85%	1,094.2	85%	536.0	100%	542.8	100%
Log-logistic	1,088.2	39%	1,095.8	39%	539.6	16%	546.4	16%
Generalised gamma	1,088.3	37%	1,099.6	6%	538.0	37%	548.2	7%
Generalised F	1,090.3	14%	1,105.4	0%	540.0	14%	553.6	0%
Gompertz	1,096.0	1%	1,103.5	1%	545.1	1%	551.9	1%
Lognormal	1,102.6	0%	1,110.2	0%	542.5	4%	549.3	4%
Exponential	1,135.6	0%	1,139.4	0%	580.8	0%	584.2	0%
Royston-Parmar models	**AIC**	**IER**	**BIC**	**IER**	**AIC**	**IER**	**BIC**	**IER**
Hazard scale; 0 knots	1,086.3	100%	1,093.9	100%	536.3	5%	543.1	100%
Hazard scale; 1 knots	1,088.2	39%	1,099.6	6%	537.9	2%	548.2	8%
Hazard scale; 2 knots	1,090.2	14%	1,105.4	0%	537.4	3%	551.0	2%
Hazard scale; 3 knots	1,091.9	6%	1,110.9	0%	533.0	28%	550.0	3%
Hazard scale; 4 knots	1,092.8	4%	1,115.6	0%	535.1	10%	555.6	0%
Hazard scale; 5 knots	1,093.7	3%	1,120.2	0%	530.4	100%	554.3	0%
Odds scale; 0 knots	1,088.2	39%	1,095.8	39%	539.6	1%	546.4	19%
Odds scale; 1 knots	1,089.2	23%	1,100.6	3%	540.7	1%	550.9	2%
Odds scale; 2 knots	1,090.8	10%	1,106.0	0%	537.1	4%	550.8	2%
Odds scale; 3 knots	1,092.4	5%	1,111.4	0%	533.5	21%	550.6	2%
Odds scale; 4 knots	1,092.8	4%	1,115.6	0%	536.2	6%	556.6	0%
Odds scale; 5 knots	1,093.8	2%	1,120.3	0%	530.9	80%	554.7	0%

AIC: Akaike's information criteria. IER: Inverse evidence ratio.

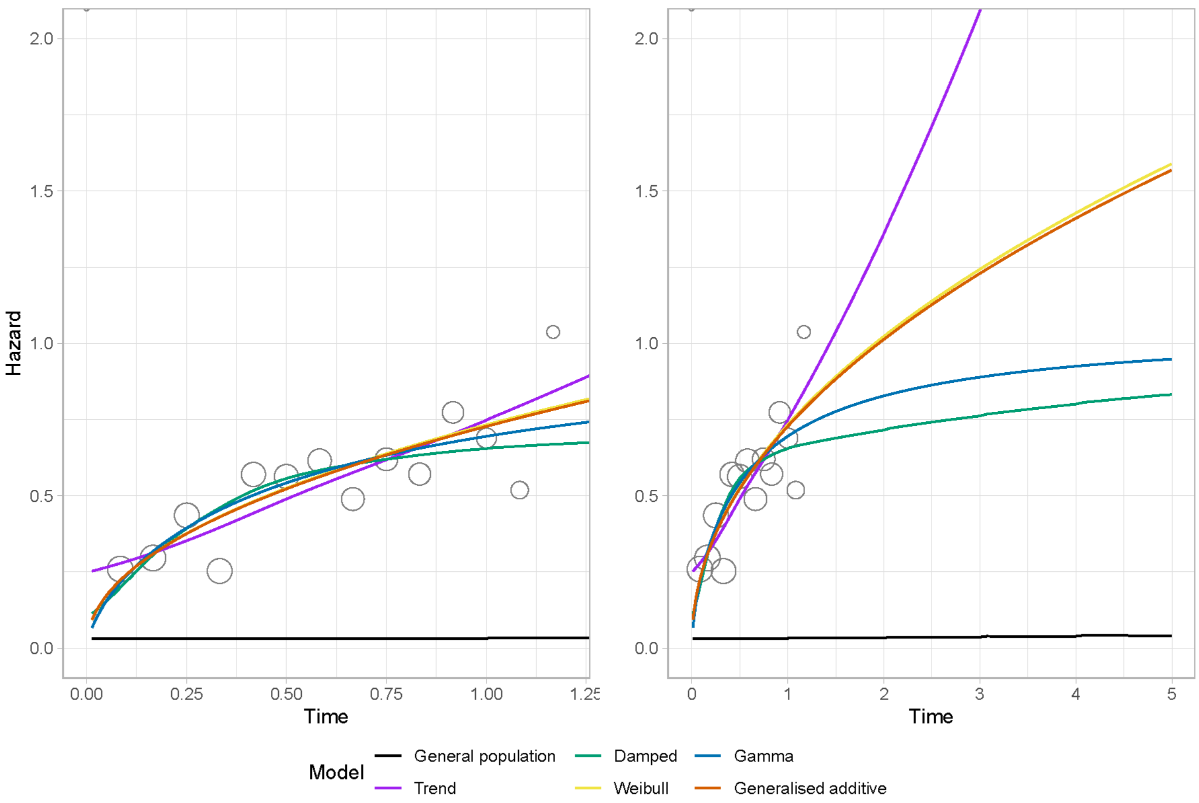
For the abiraterone group, the RPM corresponding to the Weibull (hazard scale, no internal knots) had the lowest AIC and BIC; no other model had an IER greater than 40%, with IER values decreasing with increasing model complexity. For the placebo group use of AIC and BIC led to contrary findings. The BIC supported the use of a Weibull (next largest IER was 19%, corresponding to log-logistic model), whilst more complex models had lower AIC values. Visually the more complex models appeared to be over-fitting the data, suggesting that in this instance AIC may not be sufficiently penalising model complexity. Since the best-fitting (and plausible) RPM was the Weibull for both groups and this was already chosen as a standard model, no RPMs were used for extrapolation.

The fitted GAM provided visual estimates that were very similar to the RPM with the lowest AIC, suggesting a Weibull for abiraterone and a function with six turning points in the hazard for the placebo group. These are used for extrapolation even though it is noted that the placebo GAM may be over-fitting the data (resulting in very large extrapolated hazards).

For both groups, the FP1 with the lowest AIC was the same model as a Weibull. This model provided very similar visual estimates (within-sample and extrapolations) to the best-fitting FP2 (based on AIC). For both groups the AIC for the FP1 (Weibull model) was slightly lower than that for the FP2; (placebo: 682.1 vs 683.8, abiraterone: 533.4 vs 534.5). As such, the FP1 model corresponding to the Weibull was chosen for both groups. Hence, as with the RPM, no FP models were used for extrapolation. More information on the relationship between FP(1) models and current practice models is provided in Appendix 3.

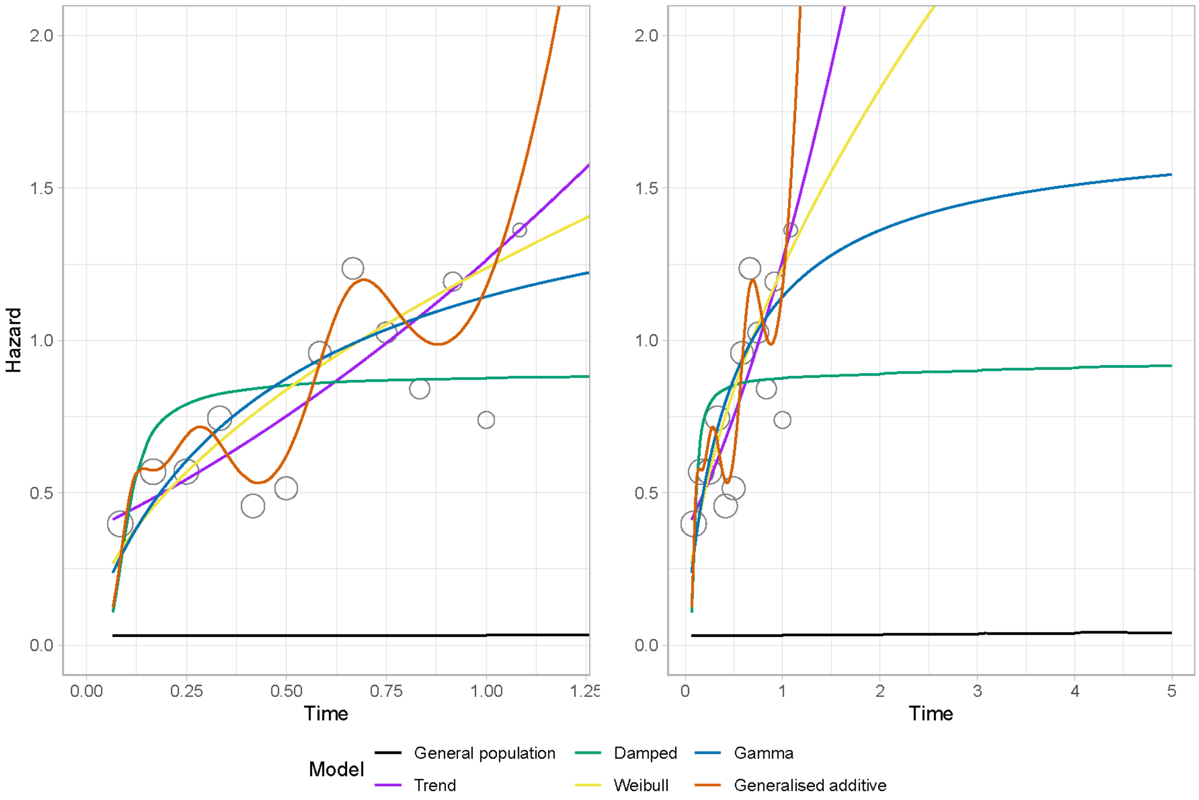
For the abiraterone group, both DSMs provide similar estimates to the Weibull model up to about 9 months. After this time, the local trend model estimates higher hazards than the Weibull and the damped trend model estimates lower hazards. Similar extrapolations were observed for the placebo group, with the damped trend providing the lowest extrapolated hazards of all models considered and the local trend the second highest (below the GAM, which appeared to be over-fitting). The higher extrapolations from the local trend model (compared with standard models) may be due to the potential increase in the hazards at the end of follow-up noted in Figure 4.1. Of all the models considered, the damped trend model is the only one which assumes that the extrapolated trend will decrease over time. The impact of this assumption is evident

Figure 4.3: Within-sample estimates and extrapolations for selected models: abiraterone.



in Figures 4.3 and 4.4, which leads to notably different extrapolations compared with the other models. Extensions to relative survival models were explored, but provided virtually identical results, so are not reported here.

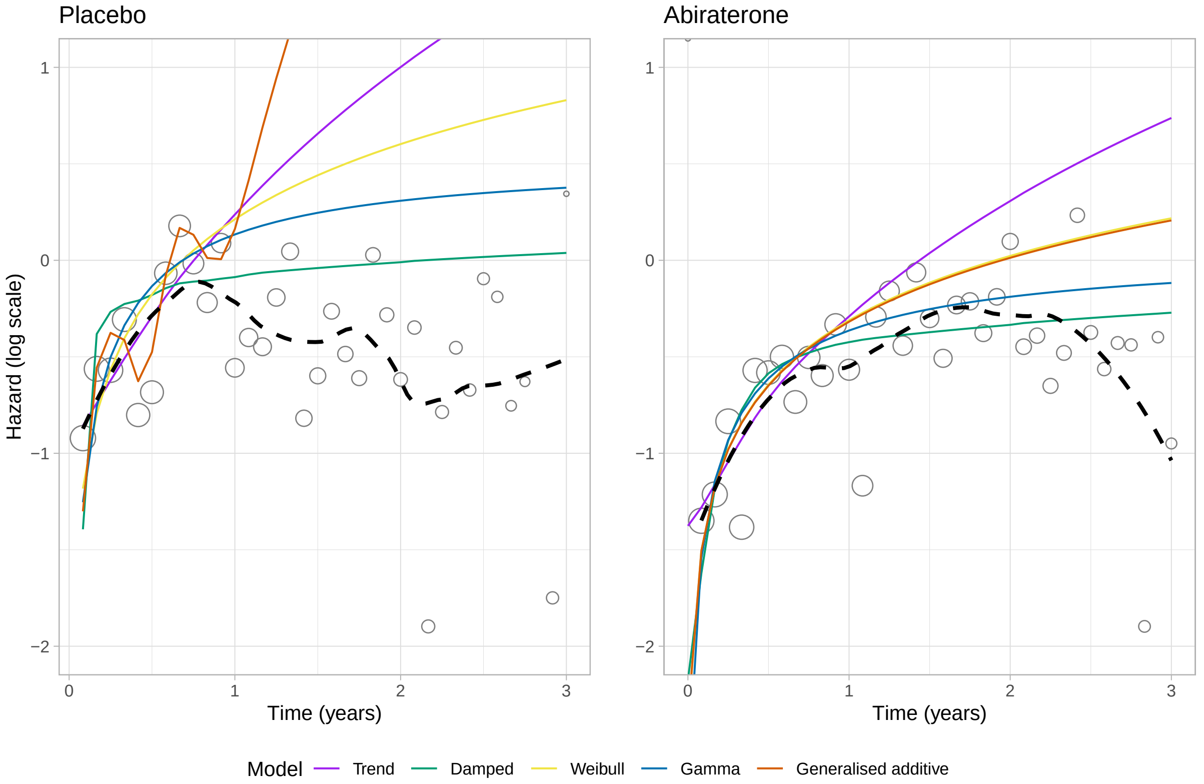
Figure 4.4: Within-sample estimates and extrapolations for selected models: placebo.



4.1.3 Goodness of fit of extrapolations

A visual comparison of the model-estimates to the longer-term data is provided in Figure 4.5, which also include a smooth non-parametric estimate (black-dashed line). For both treatment groups, the trend observed in the interim dataset does not persist in the long-term. For the placebo group, the short-term increase in the hazard during the period of the interim data is followed by an almost immediate decrease. As such, none of the models provide good extrapolations. For the abiraterone group, the hazard continues increasing to about 2.5 years albeit at a lower rate than was observed in the interim dataset. The damped trend model provides adequate extrapolations up to about 2.5 years. After this time the observed hazards decrease, and no models provide a good description. Extrapolations beyond three years were not considered due to the small sample sizes (at three years the number of patients remaining in the study was 62 and 23 for the abiraterone and placebo arms, respectively, whilst at 3.5 years the numbers were 24 and 3, respectively). Extrapolation goodness of fit measures (mean-squared error and bias, as defined in the simulation study chapters), were also calculated but are not displayed given the clear lack of fit for the extrapolations from each of the models considered.

Figure 4.5: Extrapolations from short-term data compared with longer-term data.



4.1.4 Discussion

A variety of different models were considered for extrapolating the hazard function from the interim dataset. For both treatment groups the observed hazard was increasing and the use of standard survival models favoured either the Weibull or gamma. These models both provided monotonically increasing extrapolated hazards. Similar extrapolations were obtained from the more flexible FPs, RPMs and GAMs, along with the local trend model. In contrast, the damped trend model provided extrapolations that increased at a much lower rate for both groups. All of the models predicted that hazards would continue to increase in the future. However, for both treatment groups the long-term hazards eventually decreased; this decrease occurred after approximately one and 2.5 years for the placebo and abiraterone groups, respectively. As none of the considered models were able to model a turning point in the hazard function (from increasing to decreasing) during the extrapolated period, the extrapolations were generally all poor. The damped trend model assumes that the hazard function will eventually change from increasing to constant; this is closest to what occurred in the full dataset, hence the extrapolations from the damped trend DSM were closest to the longer-term data. However, as this is a single case-study, the generalisability of this finding to other scenarios is unclear. The main finding from this case-study is that any model used for extrapolation is only as good as the dataset that it uses. If the unobserved future contains turning points, then any extrapolation model would do poorly unless it incorporates external data to identify the turning points.

4.2 Simulation study: methods

The reporting of the simulation study follows published guidance [194]. Components of the simulation study are reported based on their aims (provided in the previous section), data generating mechanisms, methods (models), estimands, and performance measures.

4.2.1 Data generating mechanism

A two-component mixture-Weibull model was used as the data generating mechanism (each component is an individual Weibull model). This model was chosen because it is capable of producing complex (multi-modal) hazard functions and may be interpreted as representing latent frailties (two sub-populations), representing patients with either a high hazard (short survival) or a low hazard (long survival) [203]. The survival and hazard functions are given by:

$$S_{t_i} = \rho \exp(-\Lambda_1 t_i^{\gamma_1}) + (1-\rho)\exp(-\Lambda_2 t_i^{\gamma_2}) \tag{4.1}$$

$$\lambda_{t_i} = \frac{\Lambda_1 \gamma_1 t_i^{\gamma_1^{-1}} \rho \exp(-\Lambda_1 t_i^{\gamma_1}) + \Lambda_2 \gamma_2 t_i^{\gamma_2^{-1}} (1-\rho)\exp(-\Lambda_2 t^{\gamma_2})}{S_{t_i}} \tag{4.2}$$

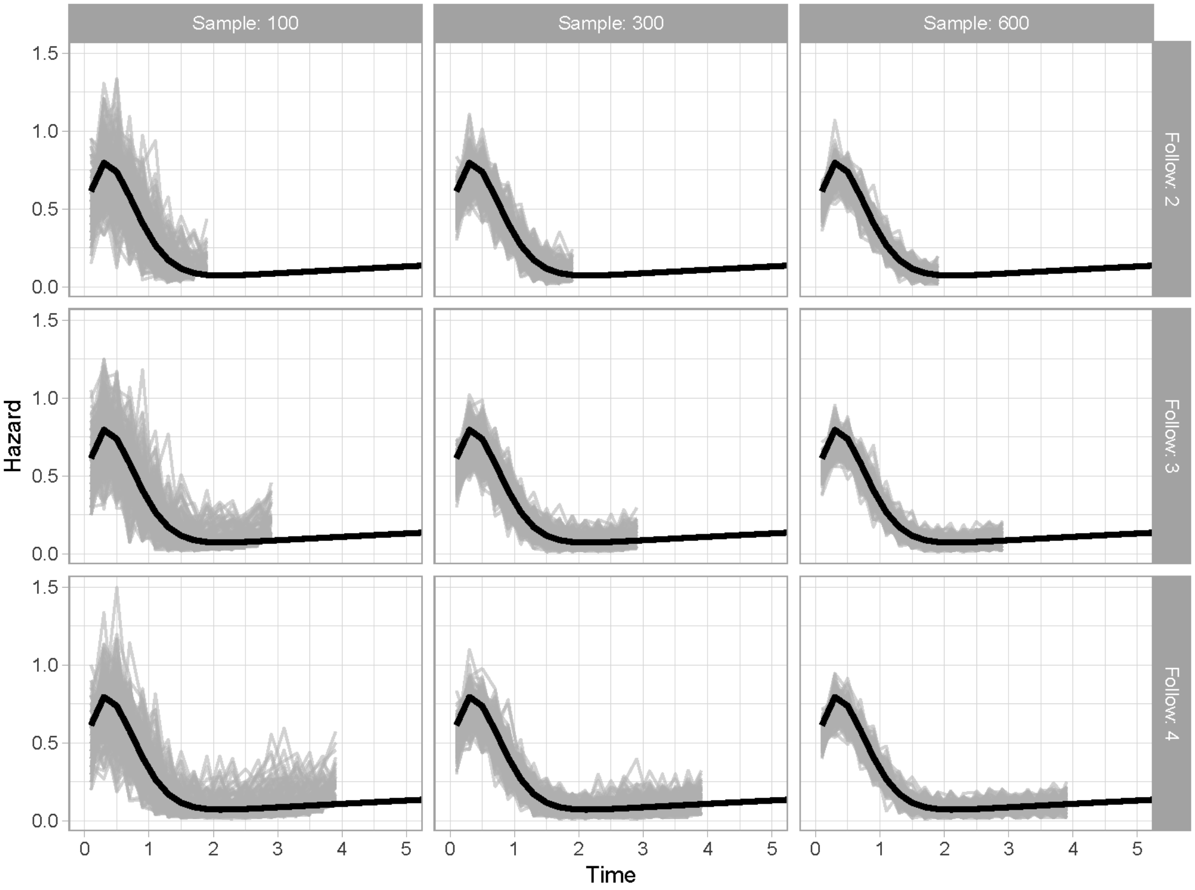
respectively, where γ and Λ are the respective shape and scale parameters (indexed by component), and ρ is the mixing proportion. The values used for this study are: $\gamma_1 = 1.8, \Lambda_1 = 0.02, \gamma_2 = 1.4, \Lambda_2 = 2.3$, and $\rho = 0.5$. A plot of λ_{t_i} is provided in Figure 4.6. The data generating mechanism was designed to reflect a ‘true’ hazard with two turning points, and a long-term increasing hazard (to reflect the impact of ageing on the hazard of death). As this is a simulation study, the time units to use are arbitrary; they were chosen to represent years. There are two turning points, at approximately 0.5 and 1.75 years. After the second turning point hazards increase monotonically.

Nine scenarios were simulated, with 200 datasets simulated for each scenario. These scenarios corresponded to three different sample sizes (small = 100, medium = 300, large = 600), and three different lengths of follow-up (short = two years, medium = three years, long = four years). The key differences between lengths of follow-up were where follow-up ended relative to the second turning point, ranging from very soon after (short follow-up) to a longer time after (long follow-up). Follow-up times less than the second turning point were not considered, as without evidence on the long-term increasing trend the competing models would not be able to provide plausible extrapolations. The sample sizes were chosen to be representative of those typically seen in clinical practice. Details on these scenarios are provided in Table 4.4. Figure 4.6 shows the simulated hazards (in grey) for each scenario along with the true hazard (in black, which is the same for each scenario). Hence, for each scenario Figure 4.6 displays 200 grey lines (hazards), where the number of observations for each hazard is equal to the sample size for that scenario.

Table 4.4: Details of the nine scenarios simulated.

Scenario	**Follow-up (survival %)**	**Sample size**
Short follow-up, small sample size	Two years (46.8 %)	100
Short follow-up, medium sample size		300
Short follow-up, large sample size		600
Medium follow-up, small sample size	Three years (43.3%)	100
Medium follow-up, medium sample size		300
Medium follow-up, large sample size		600
Long follow-up, small sample size	Four years (39.2%)	100
Long follow-up, medium sample size		300
Long follow-up, large sample size		600

Figure 4.6: Simulated hazards (grey-lines) for the nine scenarios, along with the truth (black-line)



4.2.2 Estimand and performance measures

The estimand was the mean natural logarithm of the time-varying hazard function λ_{t_i}. The natural logarithm was used as this maps λ_{t_i} to be in the range $(-\infty, \infty)$, and it may be assumed to be approximately Normally distributed. As such, both positive and negative deviations would be equally likely.

When the estimand is the mean, and positive and negative deviations from the mean are penalised equally, then the squared error is a consistent loss function (performance measure) [204]. Here consistency means that the performance measure is minimised when model estimates equal the estimand. The primary performance measure used was the mean (of the) squared error (MSE), with bias as a secondary performance measure. In addition to being a consistent loss function, the MSE has the benefit that it may be interpreted as penalising for both bias (how close are the model estimates to the truth) and variance (how much do estimates vary across simulations). Use of bias as a secondary measure provides insight into how the two components of bias and variance contribute to the MSE. The MSE and bias are defined as [194]:

$$\text{MSE}_i = \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} (\hat{\theta}_{j,i} - \theta_i)^2 \tag{4.3}$$

$$\text{Bias}_i = \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} (\hat{\theta}_{j,i} - \theta_i) \tag{4.4}$$

where n_{sim} is the number of simulations, θ_i is the estimand, and $\hat{\theta}_{j,i}$ the corresponding model-based estimate (the subscripts i, j are for time and simulation), respectively. For MSE lower values indicate better model performance, for bias values closer to zero indicate better model performance. As the hazard function is a time-varying estimand the performance measures are also time-varying. Summary (mean) values of the MSE and bias were calculated separately for the out-of-sample (extrapolations) and within-sample time periods. These summary measures use a novel method to calculate; details are provided in Appendix 1. The time-horizon was 20 years (after which survival was essentially 0%), with time-steps of 0.05 years.

4.2.3 Models

Four different classes of survival model were used: current practice, DSMs, spline-based models, and FPs. Together, the last two classes represent 'emerging practice' for this simulation study. All fitting was performed in R (v 3.5.3). For each model class, multiple model specifications were considered:

Dynamic survival models. Models with a time-varying trend were used as models without a trend would extrapolate a constant value, which was deemed to be unlikely. Both local trend and damped trend models were used; for both the level could be either global (constant) or local - this results in four models. Parameter estimation was performed using the `RStan` package [192].

Current practice. Seven models were initially evaluated: exponential, Weibull, Gompertz, lognormal, log-logistic, gamma, and generalised gamma. The main results do not include the Gompertz model, for reasons described later. Current practice models were fit using the `flexsurv` package citeRN191.

Spline-based models. Two implementations were used. One used the generalised additive models (GAMs) implemented in the `mgcv` package [135]. The default of a thin-plate spline was used, with a maximum dimension of seven and a Poisson family. The second was the Royston-Parmar model (RPM) within the `flexsurv` package. Between zero and five internal knots, on the hazard scale were considered.

Fractional polynomials. First-order and second-order models (FP(1) and FP(2) respectively) were considered. Higher-order models are possible, but rarely required due to the flexibility of FP(2) models [205]. It is possible to choose between FP(1) and FP(2) models using a closed-test procedure [205], but for this study FP(1) and FP(2) models were kept separate. This is because FP(2) models are more complex than FP(1) models, and there was interest in seeing if extrapolation performance varied by model complexity. Fitting was performed with the `stats` package.

For current practice models and the RPMs, individual patient data were used. For the DSMs, GAMs, and FP models fitting was in the GLM framework, as described in the GLM manuscript of Appendix 3, with one observed death per time interval (to avoid a loss of information). For FPs, GAMs, and DSMs it is possible to use either 'time' or log(time), with corresponding linear models being the Gompertz and Weibull. Initially both approaches were considered. Time is explicitly included as a covariate in FPs and GAMs and use of the full set of powers for FPs is only possible if the covariate is positive. To ensure this when using log(time), the actual covariate values used were log(time +1). This approach is standard for Poisson regression models where zero values are problematic and has the benefit that zero values of time are mapped to zero values of log(time) [206]. For consistency this approach was also used for GAMs. Results were worse when time was used instead of log(time) for each model class. In addition, use of the Gompertz model produced very good within-sample fits but very poor extrapolations which lacked face validity as they very quickly tended towards zero. Clinically, this is implausible since it implies that people will not die after a certain time period. As such, the results presented in this chapter use log(time). The impact of using time as the covariate is shown in Appendix 2 and the impact of including the Gompertz model in the pool of current practice models is considered in Section 4.3.1.

Multiple model specifications were considered for most of the model classes. In practice, the choice between these specifications would be based on a combination of clinical considerations (such as: are the extrapolations plausible? Does the specification reflect our understanding of the disease process?) and empirical goodness of fit (such as information criteria). As this is a simulation study, the choice between model specifications was based purely on AIC for current practice, RPM, FP(1) and FP(2) models. The AIC was chosen in preference to other information criteria as it has both theoretical support (as identifying, from a given set of models, the model that best approximates reality; see Chapter 2 of Burnham and Anderson [201]) and empirical support as a "reasonable choice" (Chapter 7 o Hyndman *et al.* [42]). Model choice was not required for the GAMs or either of the two DSMs. Of the models considered, only the DSMs are Bayesian in implementation. For these, default (uninformative) priors were used with the exception of the innovations, which were modelled by a zero-mean Normal prior for which the variance had an inverse gamma $(1, 0.005)$ prior. The choice of prior for the innovation variances was informed by the results of the *de novo* analysis in Section 3.4.

In total, nine models are presented for each scenario:

1. Local trend, global level dynamic model.
2. Local trend, local level dynamic model.
3. Damped trend, global level dynamic model.
4. Damped trend, local level dynamic model.
5. Current practice: best within-sample fit (of the six models considered).
6. Generalised additive model.
7. Royston-Parmar model: best within-sample fit (of the six models considered).

8. First-order fractional polynomial model: best within-sample fit (of the eight models considered).

9. Second-order fractional polynomial model: best within-sample fit (of the 36 models considered).

Results for each of the individual current practice models are provided in Section 4.3.1.

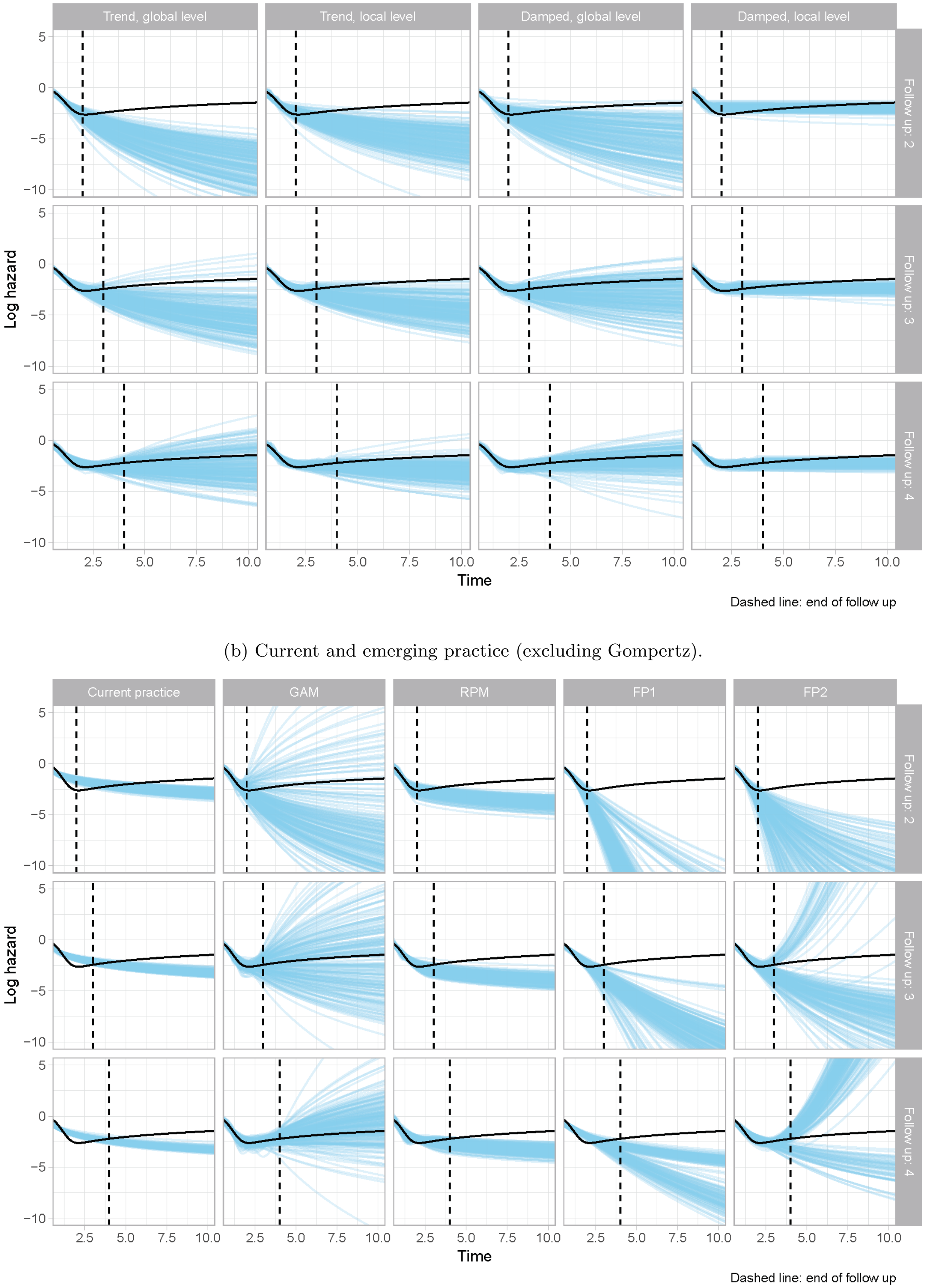
4.3 Simulation study: results

For each model, the visual patterns of within-sample fit and extrapolations were broadly similar across the nine scenarios considered. Increasing the sample size led to a reduction in the variation of extrapolations as expected but had little other effect. Results were more sensitive to changes in length of follow-up. Figure 4.7 provides results for a sample size of 300 for the three lengths of follow-up considered and all nine models. Plots for the remaining sample sizes are provided in Appendix 2. Time-varying estimates of the MSE and bias in the model estimates are provided for the three scenarios with a sample size of 300 in Figure 5.4. Figures for the remaining six scenarios are provided in Appendix 2.

For the within-sample period, the current practice models provided a poor fit to the observed data for all the scenarios: the hazard was under-estimated for the first year and then over-estimated for the subsequent years, with the turning-point in the hazard not captured. The remaining models all provided visually improved within-sample fit, although they also typically had more variability in their estimates. With the shortest follow-up (two years) none of the models identified the long-term increasing trend in the hazard function. With the longest follow-up (four years) two dynamic models (local trend and damped trend models; both with a global level) along with the GAM identified the long-term increasing hazard; the remaining models did not. For the three models that identified the long-term increasing hazard, the bias in the extrapolations decreased with increasing sample-size as they were able to identify the true trend with more accuracy. For the largest sample size the local trend (global level), damped trend (global level), and GAM all provided approximately unbiased estimates. The extrapolation performance of the two DSMs with a local level also improved with increasing sample size, but they consistently under-estimated the true hazard even with the longest follow-up. In contrast, for the current practice models and RPMs, the bias was not affected by increasing the sample size. Results for a follow-up of three years were similar to those for four years, but with more uncertainty in the extrapolations. This uncertainty led to some extreme departures from the true hazard values for the GAMs and DSMs. In contrast, use of standard models or RPMs led to extrapolations that were always biased, but there were never any extreme departures from the truth.

In general, GAMs required less data (sample size or follow-up) than DSMs to identify the turning-point in the hazard, but GAMs also produced more variable extrapolation estimates than the DSMs. This large variation is a particular concern as in a given appraisal only a single extrapolation would be observed and there is a danger that it would correspond to one of the very poor extrapolations. Both of the FP model classes provided extremely poor extrapolations; in general extrapolated hazards very quickly tended towards zero or very large numbers. In addition to being very poor, these extrapolations lacked face validity. As such, estimates from FP models are not considered further, beyond providing and commenting on summary goodness of fit values in Table 5.2.

Figure 4.7: Model estimates of the log-hazard (blue lines) and true values (black lines).



(a) Dynamic survival models.

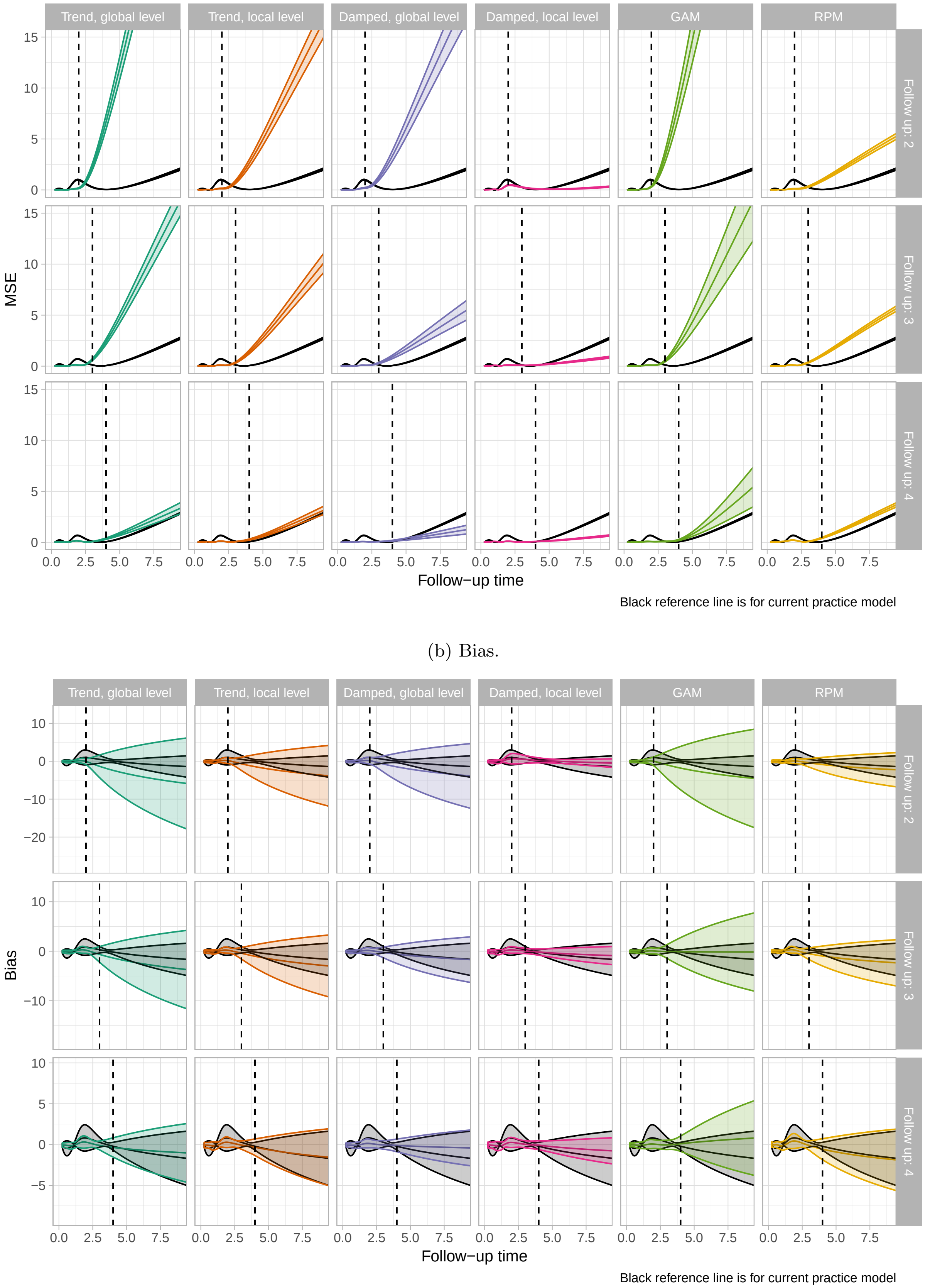
(b) Current and emerging practice (excluding Gompertz).

Despite generally having the worst within-sample fit (as measured by both MSE and bias), the current practice models often provided some of the best extrapolations with short-to-medium follow-up. However, as can be seen in Figure 4.7, the apparent good extrapolation performance of the current practice models is an artifact of their poor within-sample fit. The current practice models extrapolated the short-term decreasing trend in the hazard and failed to reflect the longer-term increasing trend. Due to the poor within-sample fit, the extrapolated (decreasing) hazards were by chance close to the true (increasing) hazards. This point is further emphasised when comparing the performance of the current practice models with the RPMs. The latter model class provided less-biased within-sample estimates for all nine scenarios, demonstrating that RPMs were better at estimating the short-term decrease in hazards. Both the current practice and RPMs extrapolated a decreasing hazard for all nine scenarios, but the RPM extrapolations were always more biased with larger MSE than the current practice extrapolations.

When comparing the four dynamic model specifications, the damped trend models produced less variation in extrapolations than the local trend models. This is to be expected, as the process of dampening is cumulative over time so that eventually extrapolations reduce to a constant value. For each of the nine scenarios the damped trend models had lower summary MSE and bias values than the corresponding local trend specification. Allowing the level to vary over time resulted in smaller trend estimates when compared to dynamic models with a fixed (global) level. This is because for global level models all of the variation in the hazard function is attributed to the trend, whilst for local level models some of the variation is subsumed by fluctuations in the level. For damped trend models with a local level, extrapolations were almost constant implying the extrapolated trend was approximately zero. The smaller trend estimates for the local level models resulted in less bias when the long-term extrapolations were incorrectly estimated to decrease (as occurred with short follow-up) but more bias when the long-term increase in the hazard function was identified (as occurred with longer follow-up). The damped trend local level model also had small variation in extrapolations, with the smallest extrapolation MSE of all nine models considered for all nine scenarios. This is despite the fact that the damped trend local level model extrapolated an essentially constant hazard for all nine scenarios.

Overall values of MSE and bias (averaged across the within-sample and out-of-sample time periods) are provided in Table 5.2. For each of the nine scenarios considered, the DSM with a damped trend and a local level provided the lowest MSE values. The next lowest MSE values were typically observed for the current practice and RPMs, despite these two model types predicting a long-term decrease in hazards for all nine scenarios. As the scenarios became more data rich (increasing follow-up and/or sample size), the performance of the DSMs improved relative to the other models. For example, with a sample size of 600 and four-years follow-up, the four DSMs had the lowest MSE of all the models considered. The class of FPs give the worst extrapolations for every scenario, as shown visually in Figure 4.7. This may be due to their sensitivity to extreme values, combined with extrapolating polynomial trends [207]. Omitting the FPs, the largest MSE values were observed for the GAM in seven of the nine scenarios. The poor performance of the GAMs is primarily driven by the large variability in extrapolations, as it provided the least-biased estimates in four scenarios. For the remaining five scenarios a DSM provided the least-biased estimates (two each for the two DSMs with a damped trend, one for the local trend global level DSM).

Figure 4.8: Mean squared error and bias values over time (within-sample and extrapolations).



(a) Mean squared error (MSE).

(b) Bias.

Table 4.5: Goodness of fit over the entire time horizon.

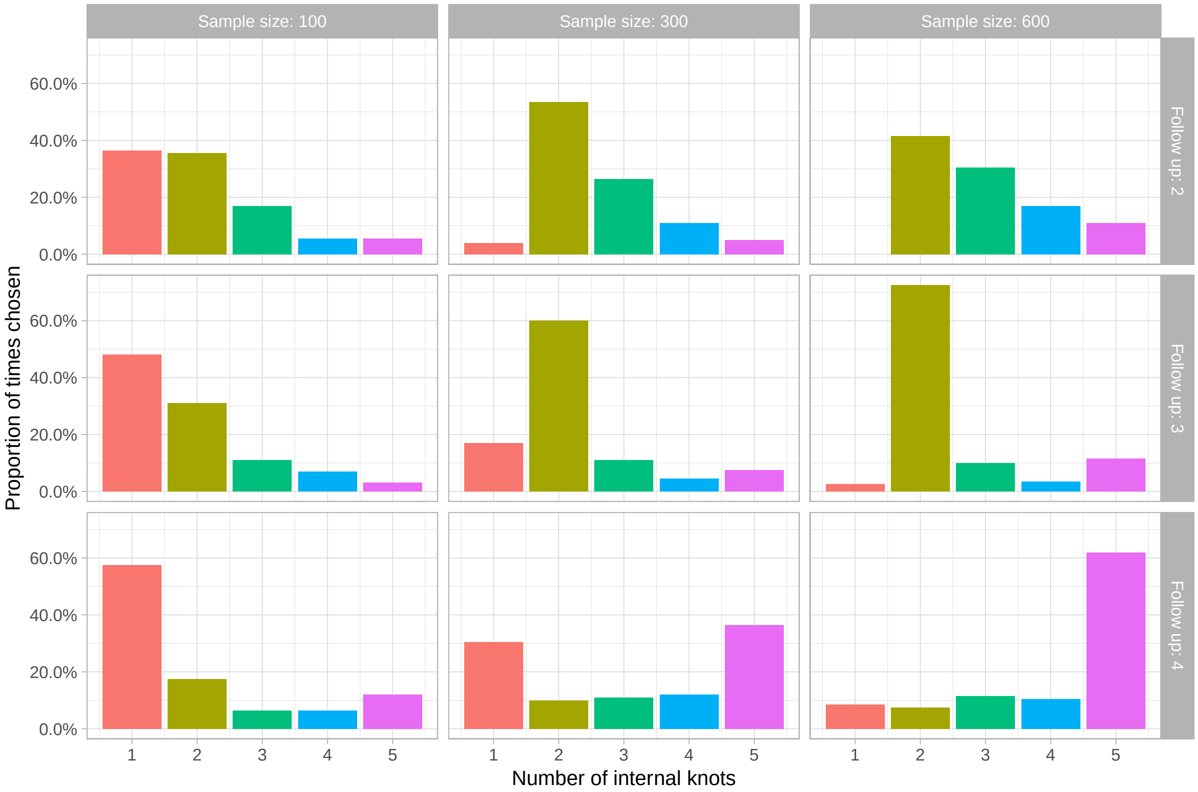
Overall mean	**Sample size 100. Follow-up:**			**Sample size 300. Follow-up:**			**Sample size 600. Follow-up:**		
squared error	**2 years**	**3 years**	**4 years**	**2 years**	**3 years**	**4 years**	**2 years**	**3 years**	**4 years**
Damped trend, local level	0.51	0.34	0.42	0.23	0.38	0.29	0.26	0.41	0.27
Current practice	1.01	1.19	1.26	0.94	1.15	1.19	0.90	1.12	1.15
Royston-Parmar model	1.98	2.38	1.87	2.21	2.38	1.50	2.25	2.36	1.40
Damped trend, global level	3.75	4.98	2.36	7.88	2.29	0.52	8.07	1.41	0.35
Local trend, local level	3.33	4.41	2.96	6.86	4.13	1.26	9.18	3.39	0.71
Local trend, global level	6.03	7.12	4.27	15.61	6.67	1.36	18.04	4.65	0.57
Generalised additive model	32.89	18.16	6.85	18.49	6.59	2.12	20.27	4.09	1.53
Fractional polynomial: order 1	312.40	103.82	22.49	326.43	41.25	8.61	331.78	35.71	9.14
Fractional polynomial: order 2	531.90	258.30	147.35	205.23	55.05	85.21	121.62	24.07	65.57
Overall bias									
Damped trend, local level	0.38	-0.03	-0.19	-0.12	-0.35	-0.40	-0.30	-0.30	-0.31
Current practice	-0.36	-0.37	-0.35	-0.55	-0.55	-0.54	-0.60	-0.58	-0.56
Royston-Parmar model	-0.92	-1.07	-1.10	-1.07	-1.10	-1.11	-0.88	-0.79	-0.77
Damped trend, global level	-0.35	-1.83	-1.80	-1.17	-0.78	-0.56	-0.76	-0.18	-0.14
Local trend, local level	-0.93	-1.80	-2.14	-1.32	-1.36	-1.23	-1.06	-0.64	-0.48
Local trend, global level	-1.36	-2.85	-3.13	-1.73	-1.72	-1.30	-1.31	-0.45	-0.11
Generalised additive model	-1.55	-2.18	-1.99	-0.06	-0.09	0.05	0.15	0.31	0.23
Fractional polynomial: order 1	-10.52	-11.87	-12.24	-5.36	-4.10	-3.91	-2.37	-1.75	-1.83
Fractional polynomial: order 2	-5.45	-8.03	-6.72	1.45	-0.79	-1.84	4.05	3.53	3.08

For both the spline-based models (RPM and GAM), model flexibility is determined by the complexity of the model: for RPMs this is the number of internal knots (varying between 0 and 5 inclusive), which was based on minimising AIC. For GAM model complexity is based on the number of parameters; as these are shrunk during parameter estimation, the result is a non-integer number (between 0 and 10 inclusive). Results for RPMs and GAMs are provided in Figure 4.9 and Table 4.6, respectively. Of note, the RPMs never chose a model with zero internal knots (which would be equivalent to a Weibull model). For the shortest available sample size, there was a preference for simple one-knot models. For the remaining sample sizes, as follow-up increased there was an increased tendency to select more complex models. For GAMs there was a strong tendency for more complex models in the scenarios with increasing sample size or follow-up. Given that the extrapolation performance of GAMs increased in these scenarios, this suggest that increasing data richness allows for models of increasing complexity. The total number of parameters in the RPMs is the number of internal knots +2. Hence, the RPMs had a tendency to select less complex models than the GAMs. For example, with a sample size of 600 and follow-up of 3 years, the majority of RPMs typically had 4 parameters, whilst the GAMs had on average 6.7 parameters.

Table 4.6: Model complexity: generalised additive models.

Sample size	**Follow up**	**Mean**	**95% confidence interval**
100	2 years	4.06	2 to 7.92
100	3 years	5.28	2 to 8.78
100	4 years	5.51	2 to 9.08
300	2 years	5.11	3.81 to 7.96
300	3 years	6.35	4.35 to 9.54
300	4 years	6.67	5.36 to 9.47
600	2 years	5.95	4.3 to 9.37
600	3 years	6.99	5.31 to 9.81
600	4 years	7.36	6.04 to 9.77

Figure 4.9: Number of internal knots chosen for Royston-Parmar models.



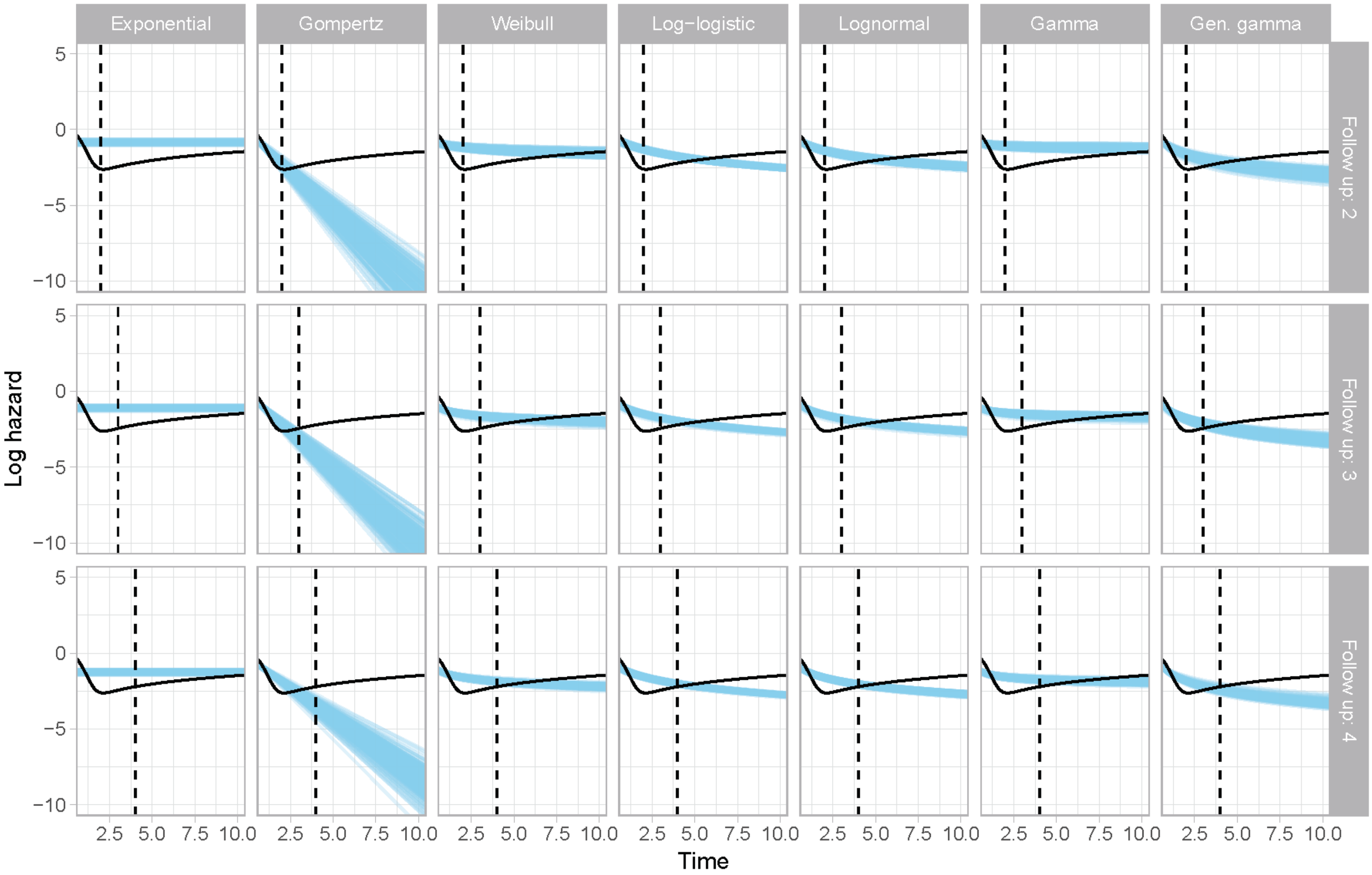
4.3.1 Current practice: the influence of the Gompertz model

As previously noted, results for current practice omit the Gompertz model as it provided very poor extrapolations, lacking face validity. This was compounded by the Gompertz model providing very good within-sample estimates, so including it would have resulted in drastically decreased extrapolation performance for the current practice models. To illustrate this issue, the goodness of fit for each of the seven models comprising current practice are provided in Figure 4.10 for the three scenarios with a sample size of 300 (results for the remaining six scenarios are near-identical and not reported). This demonstrates that none of the models are able to provide an adequate description of the long-term hazards, with all the models extrapolating a decreasing trend. Of the seven models considered, the Gompertz is the only model that does not model time on the logarithm scale (see Table 2 of the GLM manuscript in Appendix 3), but uses the original time-scale. Use of the logarithm of time compresses the time scale. For example, the difference between two and ten years follow-up changes from eight years on the original scale to 1.6 years on the logarithm scale. As such, models that use the logarithm scale for time lead to less extreme extrapolations, avoiding the implausible estimates that arise from the Gompertz model.

Summary MSE values are provided in Table 4.7, separately for the observed and extrapolated time periods. Based on within-sample fit, the Gompertz had the lowest average MSE in six of the scenarios, and the second lowest MSE (behind the generalised gamma) in the remaining scenarios. The Gompertz always had the largest out-of-sample MSE, with values that were always at least 16 times larger than the next worst extrapolating model (and more than 200 times worse than the best extrapolating current practice model in five of the scenarios). Results for bias (not shown) were similar: the Gompertz had the best within-sample and worse out-of-sample values for all nine scenarios. The best extrapolating models based on MSE (bias) were the Weibull (lognormal), gamma (Weibull), and gamma (gamma) for follow-ups of two, three, and four years respectively (irrespective of sample size). Of note, the generalised gamma nearly always provided worse extrapolations on average than its special cases (the gamma, Weibull, and lognormal). This suggests that the additional flexibility of the generalised gamma led to over-fitting temporary trends in the data.

Based on the individual simulation results, the Gompertz and generalised gamma models had the lowest within-sample MSE in 66% and 34% of simulations, respectively (no other model ever had the lowest within-sample MSE). The Gompertz never gave the best extrapolations, and the generalised gamma was only the best extrapolating model in 0.2% of simulations (4 / 1,800). The gamma was most frequently the best extrapolating model (55% of simulations), followed by the Weibull (30%) and Lognormal (12%). Whilst the Gompertz model always provided poor extrapolations for survival data, it may be a useful model in other circumstances.

Figure 4.10: Model estimates of the log-hazard (blue lines) and true values (black lines) for current practice models



Dashed line: end of follow up. Gen = generalised

Table 4.7: Summary within and out-of-sample mean-squared errors across scenarios.

	Sample size = 100		**Sample size = 300**		**Sample size = 600**	
Summary MSE values	**In-sample**	**Out-of-sample**	**In-sample**	**Out-of-sample**	**In-sample**	**Out-of-sample**
Follow-up = 2 years						
Exponential	0.90	1.41	0.88	1.39	0.88	1.40
Gompertz	0.16	47.29	0.12	46.27	0.11	45.82
Weibull	0.63	0.66	0.59	0.59	0.59	0.59
Log-logistic	0.53	0.77	0.49	0.75	0.49	0.75
Lognormal	0.46	0.67	0.43	0.66	0.43	0.65
Gamma	0.70	0.86	0.66	0.79	0.66	0.80
Generalised gamma	0.29	1.33	0.29	1.18	0.29	1.12
Follow-up = 3 years						
Exponential	1.25	0.73	1.20	0.67	1.21	0.68
Gompertz	0.18	49.02	0.15	46.96	0.14	47.32
Weibull	0.69	0.31	0.65	0.28	0.65	0.27
Log-logistic	0.61	0.78	0.57	0.78	0.57	0.78
Lognormal	0.54	0.71	0.50	0.71	0.50	0.71
Gamma	0.79	0.28	0.74	0.24	0.75	0.23
Generalised gamma	0.29	1.75	0.29	1.66	0.29	1.60
Follow-up = 4 years						
Exponential	1.11	0.35	1.09	0.33	1.09	0.33
Gompertz	0.38	36.20	0.32	33.41	0.31	33.57
Weibull	0.53	0.36	0.51	0.33	0.51	0.33
Log-logistic	0.47	0.98	0.45	0.97	0.44	0.97
Lognormal	0.42	0.91	0.40	0.90	0.40	0.90
Gamma	0.61	0.19	0.59	0.16	0.59	0.15
Generalised gamma	0.22	2.24	0.22	2.07	0.22	2.00

4.4 Simulation study: discussion

The within-sample fit and extrapolation performance of nine statistical model classes was evaluated in nine scenarios covering different lengths of follow up and different sample sizes. A single data-generating mechanism was used, with two turning points in the hazard function. Only the global-level DSMs and GAMs were able to correctly extrapolate an increasing hazard function, but only in the more data-rich scenarios, and extrapolations were highly variable. Current practice models provided the worst within-sample estimates of all the models considered. The DSMs and emerging practice models were able to provide improved within-sample fit due to their increased flexibility. However, this extra flexibility sometimes resulted in overfitting and extrapolating short term trends in the data that were not present in the longer term. A stark example of this was observed for the two FP model classes, for which extrapolations tended sharply towards implausibly small or large values. The danger of the more flexible models overfitting was in general reduced with increased sample size or follow-up, which led to improved extrapolation performance. A corresponding improvement in extrapolation performance for the more data-rich scenarios was not observed for current practice models.
Despite providing consistently biased extrapolations, current practice models exhibited very low variation in estimates across simulations. Similar results were observed for RPMs. In contrast GAMs gave the widest variation across simulations. Estimates from DSMs were less variable than from GAMs, but they required longer follow up before they could identify the increasing hazard function. Amongst the DSM specifications considered, damped trend models generally provided superior extrapolations than local trend models in all scenarios, whilst allowing the level to vary led to more conservative estimates of the extrapolated trend.

A strength of this simulation study is the large number of survival models considered. For each scenario DSMs, current practice, spline-based models, and fractional polynomials were all evaluated. When including different model specifications, collectively 62 different models were fit for each scenario, with nine models retained for estimating extrapolation performance. In the sensitivity analysis that used models as a function of time a further 47 models were considered. The large number of models included in this simulation study provides important insights into how extrapolation performance varies with model specification. For example, current practice

models provided poor within-sample estimates, and their extrapolation performance remained biased with little effect of increasing follow-up. In contrast DSMs and GAMs provided very good within-sample fits, with increasing extrapolation performance as the follow-up time increased. Another important finding was that the use of log-time in the specification of models led to superior extrapolation performance than use of time for all of the models considered.
The use of model selection also showed that within-sample goodness of fit plays a very limited role in identifying models that provide accurate extrapolations. For example, the current practice model with the best within-sample fit typically provided the worst extrapolations. A further strength of the study is the use of time-varying estimands instead of a single summary measure of accuracy such as the estimate of lifetime mean survival, which is affected by both within and out-of sample fit (an accurate estimate may occur if short-term over-estimates of hazard and long-term under-estimates cancel out, or vice-versa).

To this author's knowledge the extrapolation performance of DSMs, GAMs and FPs has not been assessed before, in either simulation studies or case-studies. Neither have previous studies considered the impact of modelling on the time scale versus using the logarithm of time. Hence the findings of this study provide valuable contributions to the literature. The use of time-varying estimands, and the methods to combine these into a summary measure, also represent a novel development.

There are limitations to this work. Only a single data generating mechanism (a mixture Weibull) was considered, with only one set of parameters. The within-sample fit and extrapolation performance of the candidate models in other settings is currently unknown and would be a fruitful area for future research. The existing data generating mechanism included two turning points, so in this sense favoured the more flexible models. However, survival data are inherently complex with a multitude of potential competing effects, such as ageing, frailty, treatment benefits, and adverse events. Collectively these are likely to cause complex shapes in the hazard function. As this is a simulation study, the generalisability of the results depends on the simulated data being realistic. The simulated data had a short-term increase in the hazard function, followed by a decrease then a longer-term increase. For this thesis three case-studies in extrapolating overall survival are presented (one each in Section 4.1 and Chapter 6, and one in the GLM manuscript of Appendix 3). These all have a short-term increase followed by a decrease in the hazard function; it may be argued that in the long-term the hazard function is likely to increase again due to the impact of ageing. As such, this suggests that the data generating mechanism used in this analysis is not unrealistic, and so the results will be useful to general practice.

In this study a total of (9 × 200 =) 1,800 datasets were simulated and for each 62 models were fit, resulting in 111,600 sets of extrapolations. Because of this, model selection was largely automated and based on minimising AIC where multiple model specifications were available (such as for the 6 RPMs). The one exception is for current practice models, where the Gompertz was omitted on the grounds of its extrapolations lacking clinical plausibility. This does not fully reflect the extrapolation task; it is likely that further improvements in the extrapolation performance of the flexible models would have been observed if clinical plausibility were used to adjust model specification when their extrapolations were clinically implausible. This is likely to particularly affect results for GAMs and global level DSMs, as extrapolations from these showed large variability.

Given that several models provided biased extrapolations (such as the current practice and RPMs for all nine scenarios), it was felt that exploring the accuracy of estimates of uncertainty would be of very limited value. Accurately quantifying uncertainty is important, but secondary to the task of providing accurate extrapolations. Estimates of uncertainty that arise from the use of different model classes is beyond the scope of this thesis, but it is briefly explored in Section 6.4.1.

In addition to considering the performance of these models under alternative data generating mechanisms, further simulation studies could seek to identify if there are certain situations when one or more of the model classes out-performs the other models, and so may be used as the default approach. The current results suggest that whilst use of a DSM may be beneficial, there is a danger that it will provide worse extrapolations than current practice models, especially in data-poor scenarios. This suggests the use of a variety of different model classes in scenario analysis, with the choice of base-case model made on a case-by-case basis. This choice would consider the specifics of the extrapolation problem, such as the plausibility of extrapolations, the richness of the available data, the qualitative differences in extrapolations arising from different models, and how extrapolations compare to any external evidence that is available.

The class of spline models considered in this simulation study included both RPMs and GAMs. These provided similar within-sample estimates but markedly different extrapolations; RPMs provided stable extrapolations, but failed to identify the long-term increasing hazard function, and their extrapolation performance did not improve with increasing follow-up or sample size. In contrast, extrapolations from GAMs were highly variable but were often able to correctly identify the increasing hazard trend, and extrapolation performance improved in more data-rich scenarios. Key differences between RPMs and GAMs are that the former use knot placement for cubic polynomials with post-hoc correction for model complexity, whilst the GAMs used here employed thin-plate regression splines, did not require the specification of knots, and model complexity was penalised during parameter estimation. Future research could explore which of the differences in model specification contribute to differences in extrapolation performance, and for both GAMs and RPMs if alternative model specifications would further improve performance. For example, an existing case-study noted that the extrapolation performance of RPMs can be sensitive to the choice of knot location [208]; default values (placed at quantiles of the observed death times) were used in this simulation study.

4.5 Conclusions

The results of the abiraterone case study illustrate an important point: without incorporating external evidence, survival models will only provide accurate extrapolations if the data to which they are fit is predictive of the future. This point is in some sense obvious, but it is important to emphasise, as without long-term evidence it is not possible to validate the assumptions used when extrapolating. For the simulation study, all nine scenarios included the last turning-point in the hazard function within the observed period of follow-up. If shorter follow-up times had been used, extrapolation performance would have been worse. The case study has supplemented the simulation study of this chapter by considering the extrapolation performance of dynamic models when using a real dataset of observed survival.

The simulation study demonstrated that in situations when survival outcomes may arise from distinct patient populations, current practice models are unlikely to provide accurate estimates of the observed data or realistic extrapolations. Of the alternative models considered, DSMs and GAMs were the only ones able to capture the long-term behaviour of the hazard function. However, extrapolations from these more flexible models were more variable than extrapolations from current practice models and had the potential to be less accurate. Of the flexible models, GAMs were able to estimate the long-term behaviour with less mature data than DSMs, but they also provided more variation in estimates.

The data generating mechanism used in the simulation study was a mixture Weibull model. This reflects two distinct patient populations, with each experiencing a different hazard of mortality. A related, but conceptually distinct, situation arises when one of the patient populations may be viewed as cured, and so experiences only general population mortality. This situation allows for the incorporation of external evidence in the form of life tables to describe general population mortality. The extrapolation performance of DSMs and comparator methods in this situation is considered in the next chapter.

Chapter 5

Comparing cure fraction models

Within the HTA literature there has been growing interest in the assessment of health technologies with a potentially curative effect [134, 208, 209, 210]. A characteristic of these treatments is that a sub-group of the treated population may be viewed as 'cured' from their disease, and so experience survival that is the same as the corresponding general population. The overall average hazard function that is observed is then a mixture of two hazard functions; a disease-specific hazard and a general population hazard. Parametric survival models have been developed to reflect the presence of a cured fraction (as identified in the methodological mapping review of Chapter 2), and for this thesis a novel extension of DSMs to incorporate a cured fraction (giving DCFMs), is described in Chapter 3. These models incorporate external data in the form of general population life tables. The extrapolation performance of these 'cure models' was assessed in a simulation study, reported in this chapter. Extensions of current practice models to incorporate a cure fraction have been used in HTA but estimates of the cure fraction have been shown to be very sensitive to model specification. For example, in a NICE appraisal estimates of the cure fraction varied by approximately 35% depending on the cure model used, whilst in a case-study it varied from 0% to 23%. [208, 211]. To assess the impact of not incorporating external data, models which do not assume a cure fraction were also considered, as was the impact of applying cure models to data that were simulated from a data-generating mechanism without a cure fraction (the mixture-Weibull of Chapter 4).

5.1 Methods

The reporting of the simulation study methods uses the same guidance and structure as Chapter 4, and so covers aims, data generating mechanisms, estimands and performance measures, and models [194]. These are discussed in turn.

5.1.1 Aims

This simulation study had two primary aims.

1. To compare the goodness of fit, for both within-sample and extrapolations, of models with a cure fraction. This includes dynamic models, models representing current practice, and models representing emerging practice.
2. To assess the impact of not incorporating external data, by comparing extrapolations from cure models against extrapolations from models which do not assume a cure fraction.

There were three secondary aims.

1. To identify the accuracy with which the true cure fraction was estimated.
2. To evaluate the performance of the cure models when the true data did not include a cure fraction.
3. To explore the effect of incorrectly including external evidence (in the form of relative survival models) in the two situations when the truth does and does not include a cure fraction.

5.1.2 Data-generating mechanisms

The ‘true’ survival and hazard functions were simulated from a Weibull cure data-generating mechanism. With this mechanism, hazard and survival functions are represented by a Weibull model for individuals who will die from the disease. For individuals who are cured from the disease general population life-tables are used to estimate the hazard and survival functions. A

Weibull model with shape (γ) and scale (Λ) values of 1.6 and 2.6, respectively was used. This provides a mean disease-specific survival of 2.33 years, with survival less than 0.05% at eight years. Data on general population (background) hazard and survival were taken from the English 2016 life-tables [212] assuming a starting age of 63 years (the mean age in the case-study of Chapter 6). It was assumed that one quarter of the sample would be cured ($\rho = 0.25$).

The equations for obtaining the observed survivor and hazard functions were provided in Equation (Chapter 3), and are replicated here:

$$
\begin{aligned}
S_{t_i} &= S^P_{t_i}, \times \rho + S^U_{t_i} \times (1 - \rho) \\
\lambda_{t_i} &= \frac{\lambda^P_{t_i} \times S^P_{t_i}, \times \rho + \lambda^U_{t_i} \times S^U_{t_i} \times (1 - \rho)}{S_{t_i}}
\end{aligned}
$$

The time-units are interpreted as years. Details of the individual components are provided in Figure 5.1. The disease-specific and background hazards are both monotone-increasing, with larger hazards observed for the disease-specific group, as expected. The observed hazard is a (time-varying) mixture of the two hazards, where the weights are the corresponding proportions of people still alive. The observed hazard function has two turning points. There is an initial increase in the hazard driven primarily by the strong increase in hazard for the uncured group. Over time, as the proportion of the alive sample who are uncured decreases, the rate of increase in the hazard slows down. After about 2.5 years the overall hazard changes from increasing to decreasing, as it approaches the general population hazard function. This is reached after about seven years (when almost all the uncured individuals are dead); after which the hazard function again increases.

As with the simulation study of Chapter 4, nine scenarios were considered, with 200 datasets simulated for each scenario. These scenarios corresponded to three different sample sizes (small = 100, medium = 300, large = 600), and three different lengths of follow-up (short = two years, medium = four years, long = eight years). When defining length of follow-up, the longest follow-up was chosen so that there were almost no uncured individuals still alive. The shortest follow-up was chosen to be representative of the lengths of follow-up often seen in cancer trials [63], with medium follow-up falling in-between. Of note, the shortest follow-up scenarios do not include the initial turning-point in the hazard, whilst both the medium and long follow-scenarios include the first but the second turning-point is only included in the long follow-up scenarios. The sample sizes were chosen to represent typical sample sizes of cancer treatments. Details on the nine scenarios are provided in Table 5.1. Figure 5.2 shows the simulated hazards for each scenario (in grey) along with the true hazard (in black, which is the same for each scenario).

Figure 5.1: Hazard and survival values for the cured (general population) and uncured (disease-specific) groups, with the average (observed) values

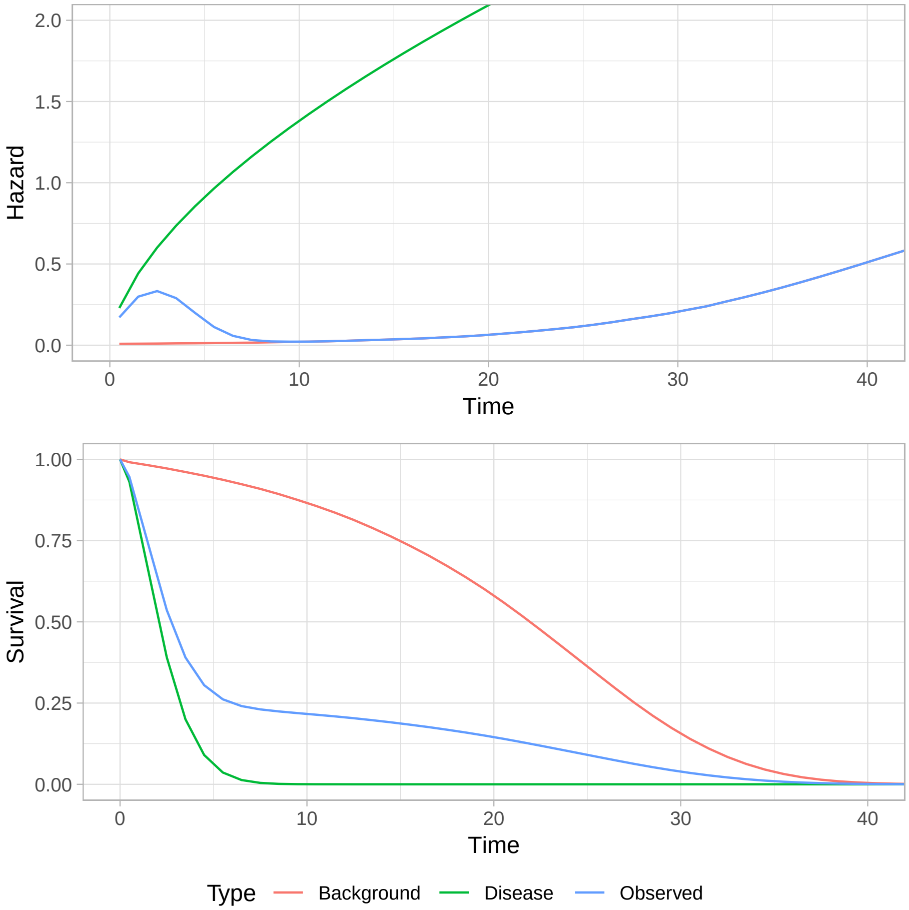
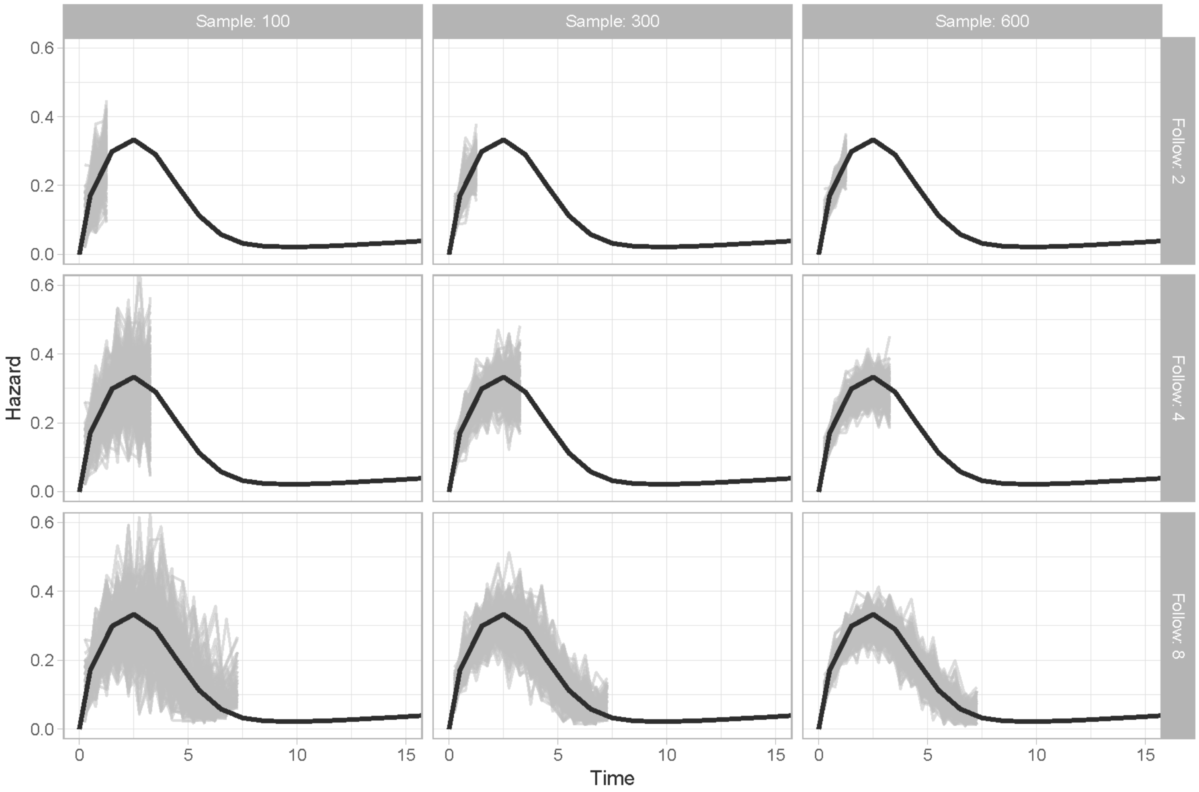


Table 5.1: Details of the nine scenarios simulated.

Scenario	**Follow-up (survival %)**	**Sample size**
Short follow-up, small sample size	2 years (63.9%)	100
Short follow-up, medium sample size	Cured = 97.7%	300
Short follow-up, large sample size	Uncured = 52.6%	600
Medium follow-up, small sample size	4 years (34.8%)	100
Short follow-up, medium sample size	Cured = 95.5%	300
Short follow-up, large sample size	Uncured = 14.5%	600
Long follow-up, small sample size	8 years (22.7%)	100
Short follow-up, medium sample size	Cured = 90.1%	300
Short follow-up, large sample size	Uncured = 0.3%	600

Figure 5.2: Simulated hazards (grey-lines) for the nine scenarios, along with the truth (black-line)



5.1.3 Estimand and performance measures

The estimand was the natural logarithm of the time-varying hazard function, with primary and secondary performance measures of MSE and bias, respectively. The estimand and performance measures are the same as those used in the simulation study of Chapter 4, which provides further details on these and justification for their use. A time-horizon of 40 years was used (at which overall survival was 0.1%), with time-steps of 0.05 years.

5.1.4 Models

A key distinction between the models considered was if external evidence (general population mortality) was included or not. These situations are discussed in turn.

Models with external information; cure models

The following models were considered. All the models are able to incorporate external evidence and make the assumption that a proportion of the sample will be cured. For all the models the correct background mortality was supplied for the cured group. Hence the models estimated the cure proportion and the hazard function for the uncured group.

Weibull cure model. This model is the same as the data-generating mechanism, so the model structure is 'correctly specified'.

Lognormal cure model. In contrast to the previous model, the model structure is now 'incorrectly specified' (misspecified) with regards to the true functional form of the hazard for the uncured group.

Royston Parmar cure model. Between zero and four internal knots were considered. Models were fit on the hazard scale, so include as a special case the Weibull cure model (when there are zero knots). As such, the model is 'over specified' - it includes the correctly specified model as a special case but allows for a wider range of more flexible models. For each dataset, the model with the lowest AIC was used to generate extrapolations.

Dynamic cured fraction models. Two models were considered: a local trend model, and a damped trend model. Both models had a global level. The local trend model is over-specified (it is the same as the Weibull cure model if the innovation variance = 0), whilst the damped trend model is incorrectly specified.

Of the current practice models, cure fraction versions are only implemented in existing software for the exponential, Weibull and lognormal. The exponential was not considered as it was assumed that there would be sufficient subject-matter knowledge to rule-out a constant hazard for the uncured fraction. Extensions to model a cure fraction are not currently available for fractional polynomials or GAMs, hence these models were not considered. Based on the results of the simulation study in Chapter 4 the extrapolation performance of GAMs was broadly similar to that of DSMs, whilst the performance of FPs was generally very poor. As such, their omission is not expected to majorly limit the conclusions drawn. The range of models considered allows for both a comparison of DCFMs against other cure models, and an exploration of the impact of model misspecification or over-specification. All model fitting was done in R. DCFMs were fit using `RStan`, the remaining models were fit using the `cuRe` package [192, 140]. The specification of the DCFMs was the same as the DSMs as described in Chapter 4, except with an additional cure fraction. The proportion of excess deaths observed during trial follow-up (based on comparing observed deaths with those expected from general population mortality) was removed from the upper-bound of the cure fraction as it was assumed that these deaths occurred amongst uncured individuals. The lower bound for the cure fraction was zero, and it had a uniform prior distribution amongst these bounds.

Models with external information; relative survival models

Dynamic relative survival models (DRSMs) were considered. These incorporate external evidence on the general population hazard function under the assumption that it provides a lower bound for the overall hazard function. As with the DCFMs, two specifications were evaluated, with either a local trend or a damped trend model (both with a global level), with fitting using the `RStan` package. Further details on model specification are provided in Section 3.3.

Models without external information

The models without external information were the same as those used for the simulation study of Chapter 4, which provides a detailed description. Again, where multiple model specifications were possible, model choice was based on minimising AIC, to provide an automated method that reflects current approaches to model choice [20]. Key details are repeated below.

Standard practice. Six models were evaluated: exponential, Weibull, lognormal, log-logistic, gamma, and generalised gamma. One model was retained. The Gompertz was omitted as it provided implausible extrapolations for follow-up times of two and four years: extrapolations quickly tended towards infinity, which lacked face validity given the observed range of hazard values. In a separate analysis (reported in Section 5.2.2) the individual results for these models are considered, along with the Gompertz.

Fractional polynomials. First-order and second-order models (FP(1) and FP(2) respectively) were considered. One model was retained for each of the FP(1) and FP(2) models.

Spline-based models. Two implementations were evaluated. One used the generalised additive models (GAMs) implemented in the `mgcv` package. The second was the Royston-Parmar model (RPM), with up to five internal knots (no internal knots being the same as the Weibull model), as estimated by the `flexsurv` package, with one RPM retained.

Dynamic survival models. Two models were used: a local trend DSM and a damped trend DSM. Both models had a global level.

This provided seven models for which the within-sample and extrapolation performance were examined.

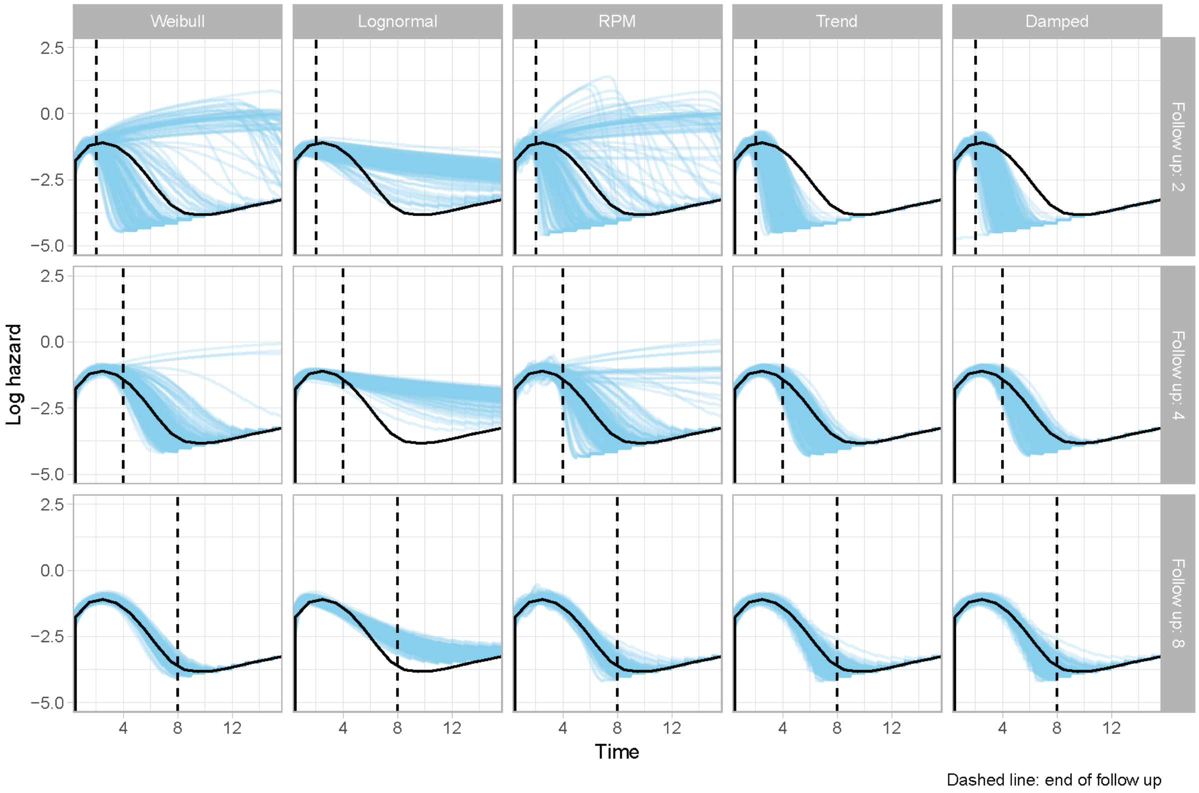
5.2 Results

For the nine scenarios considered, varying the length of follow-up had a larger impact on results than varying the sample size. Hence the results presented here are for the three follow-ups (two, four, eight years), with a sample size of 300. Results for the remaining six scenarios are provided in Appendix 2.

5.2.1 Cure models

A visual comparison of the model extrapolations and the true log-hazards is provided in Figure 5.3. With the shortest follow-up, all the considered models provided poor predictions. The Weibull and RP cure models both resulted in very similar extrapolations, with both exhibiting the largest variation of the five models considered. Extrapolations from the Weibull and RP cure models generally fell into one of two categories. For the first a large, persistent over-estimation of the hazard function arose in models which estimated a cure fraction of approximately zero. For the second the cure fraction was over-estimated, resulting in an under-estimate of the true hazards (extrapolations quickly fell to general population hazards), until the true values also matched the general population hazards. Extrapolations from the DCFMs also under-estimated the true hazards and were visually very similar to those from the second category of Weibull and RP cure models extrapolations. The lognormal cure model provided extrapolations that nearly always over-estimated the truth. These extrapolations were also visually distinct to the first category of extrapolations from the Weibull and RP cure models due to the different model structure used. With the shortest follow-up, increasing the sample size led to less variation in extrapolations from the lognormal cure model (but not improvement in fit) and had a negligible impact on extrapolations from the remaining four cure models. For all five cure models visual goodness of fit improved as the length of follow-up increased; with the longest follow-up all the cure models provided very good fits except for the lognormal cure model, which continued to systematically over-estimate the truth.

Figure 5.3: Estimates of the log-hazard compared to the truth: cure models



Estimates of bias and MSE over time are provided in Figure 5.4 whilst summary measures of MSE and bias averaged over the entire time horizon are provided for all nine scenarios in Table 5.2. Results for the remaining six scenarios are provided in Appendix 2. Within-sample estimates were all unbiased (95% confidence intervals all included zero) with point-estimates very close to zero. Across the nine scenarios, mean within-sample bias ranged from -0.09 to zero (both DCFMs), -0.01 to 0.04 (RP cure model), -0.03 to 0.04 (Weibull cure model) and -0.06 to 0.10 (lognormal cure model). For a given sample size, the worst within-sample bias estimates were always for the longest follow-up. This will be because the longest follow-up represents the most complex observed hazard function. Similar within-sample results were obtained for MSE values which, when compared to the range of MSE values for the extrapolated period, were very minor.

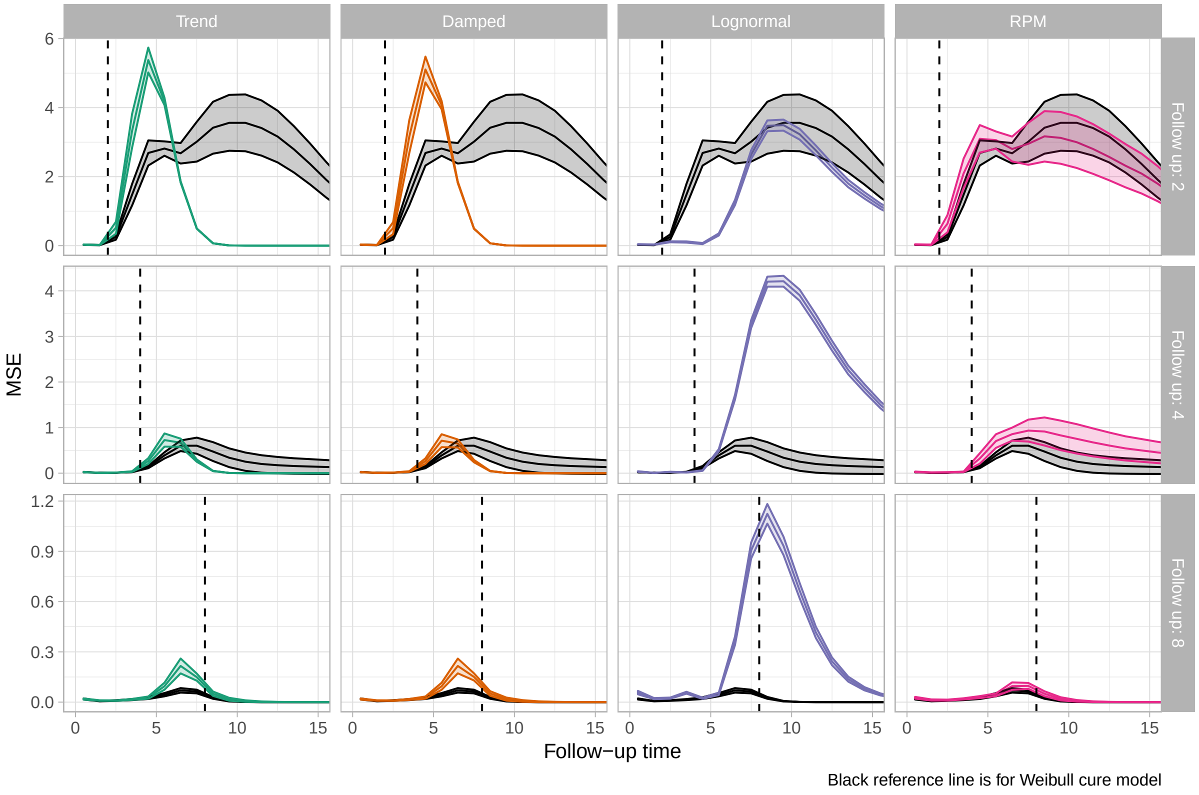
In the scenarios with the longest follow-up, all the models apart from the lognormal had overall (within-sample and extrapolated period) MSE and bias values that were very close to zero.

Further, except for the lognormal, the goodness of fit of the models generally increased as the length of follow-up or the sample size increased. The lognormal cure model had similar within-sample fit to the other models; the poor overall performance of this model was due it providing poor extrapolations.

As noted for the visual fit, bias across the entire time horizon is largest in the scenarios with the shortest follow-up, and approximately zero with the longest follow-up. The exception to this is the lognormal cure model, which remains biased even with eight years follow-up. When compared against the correctly specified model (Weibull cure model), the lognormal cure model always had estimates of bias that were both larger and more variable, whilst the RP cure model provided almost identical point-estimates but larger variation. With short and medium follow-up, the two DCFMs both had more bias than the Weibull cure model with more variation for short-term extrapolations, but in the longer-term provided unbiased estimates with less variation than the Weibull cure model. Results for MSE were generally the same as for bias, except for the lognormal for the shortest follow-up, as this had generally lower MSE than the Weibull cure model. This low MSE arises because extrapolations for the lognormal cure model, despite being more biased than the Weibull cure model, had much lower variance. Varying the sample size had very little impact on the results. Hence, for the non-dynamic cure models, model misspecification (using a lognormal cure model instead of a Weibull cure model) led to an increase in the bias of extrapolations, whilst using an over-specified model (RP cure model) had a negligible impact on bias. The overall goodness of fit of the DCFMs was similar to that of the correctly specified Weibull cure model, despite being over-specified (local trend) or misspecified (damped trend). Based on overall goodness of fit, with shorter follow-up the two DCFMs typically provided the best MSE values. Overall, the Weibull cure model and damped trend cure models had the lowest MSE in four scenarios each. The remaining scenario had the least mature data (follow-up two years and sample size = 100). For this, the lognormal cure model provided the best MSE. The second lowest MSE was from a DCFM in six scenarios, and the RP cure model in the remaining three.

Figure 5.4: Mean squared error and bias values over time (within-sample and extrapolations). Sample size = 300.

(a) Mean squared error (MSE).



(b) Bias.

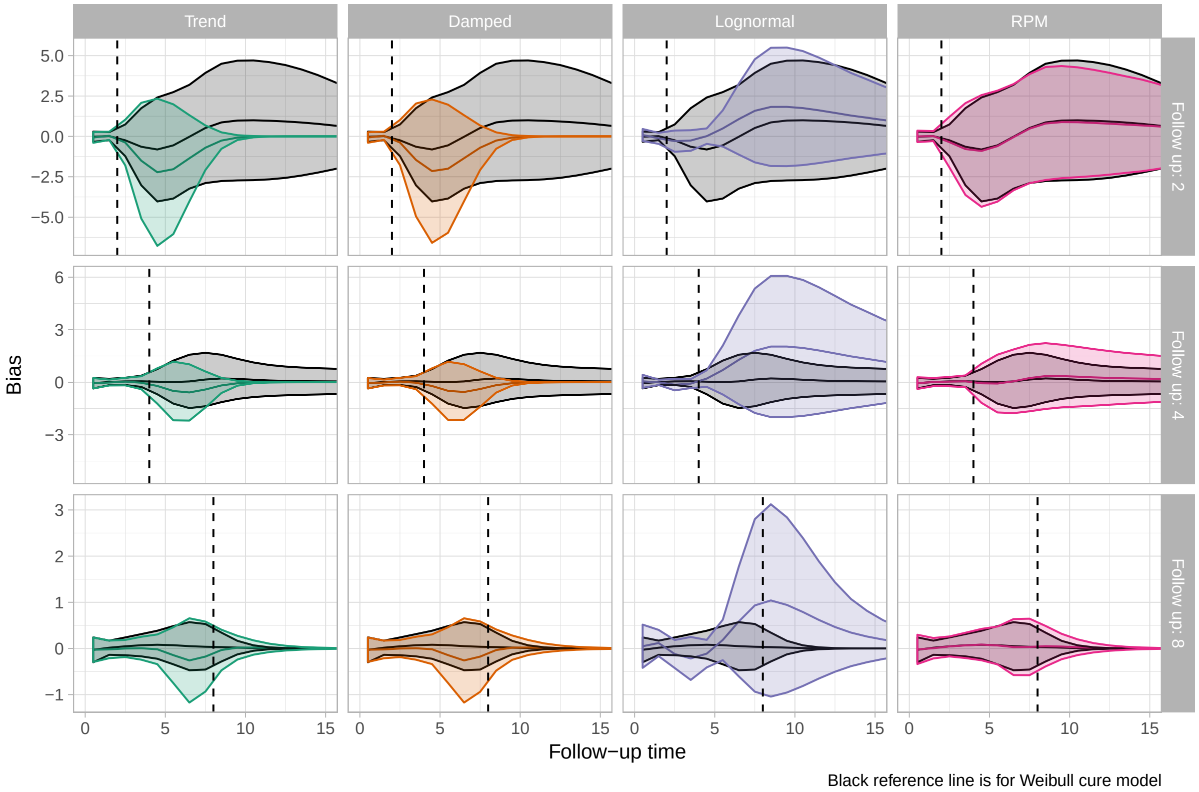


Table 5.2: Goodness of fit over the entire time horizon: cure fraction models.

Overall mean	**Sample size 100. Follow-up:**			**Sample size 300. Follow-up:**			**Sample size 600. Follow-up:**		
squared error	**2 years**	**3 years**	**4 years**	**2 years**	**3 years**	**4 years**	**2 years**	**3 years**	**4 years**
Weibull cure model	1.73	0.34	0.04	1.35	0.14	0.01	0.84	0.04	0.01
Lognormal cure model	0.85	0.87	0.23	0.87	1.04	0.17	0.92	1.06	0.17
Royston-Parmar cure model	1.77	0.34	0.06	1.36	0.31	0.02	0.91	0.22	0.01
Local trend cure model	0.90	0.13	0.07	0.70	0.08	0.02	0.66	0.06	0.01
Damped trend cure model	0.87	0.13	0.07	0.68	0.08	0.02	0.64	0.06	0.01
Overall bias									
Weibull cure model	0.20	0.16	0.08	0.07	0.04	0.01	0.01	0.02	0.02
Lognormal cure model	0.33	0.45	0.48	0.49	0.59	0.60	0.18	0.20	0.21
Royston-Parmar cure model	0.08	0.12	0.08	0.05	0.10	0.07	0.01	0.02	0.02
Local trend cure model	-0.42	-0.37	-0.37	-0.10	-0.08	-0.07	-0.04	-0.03	-0.02
Damped trend cure model	-0.42	-0.37	-0.36	-0.10	-0.08	-0.06	-0.04	-0.03	-0.02

Estimate of the cure fraction

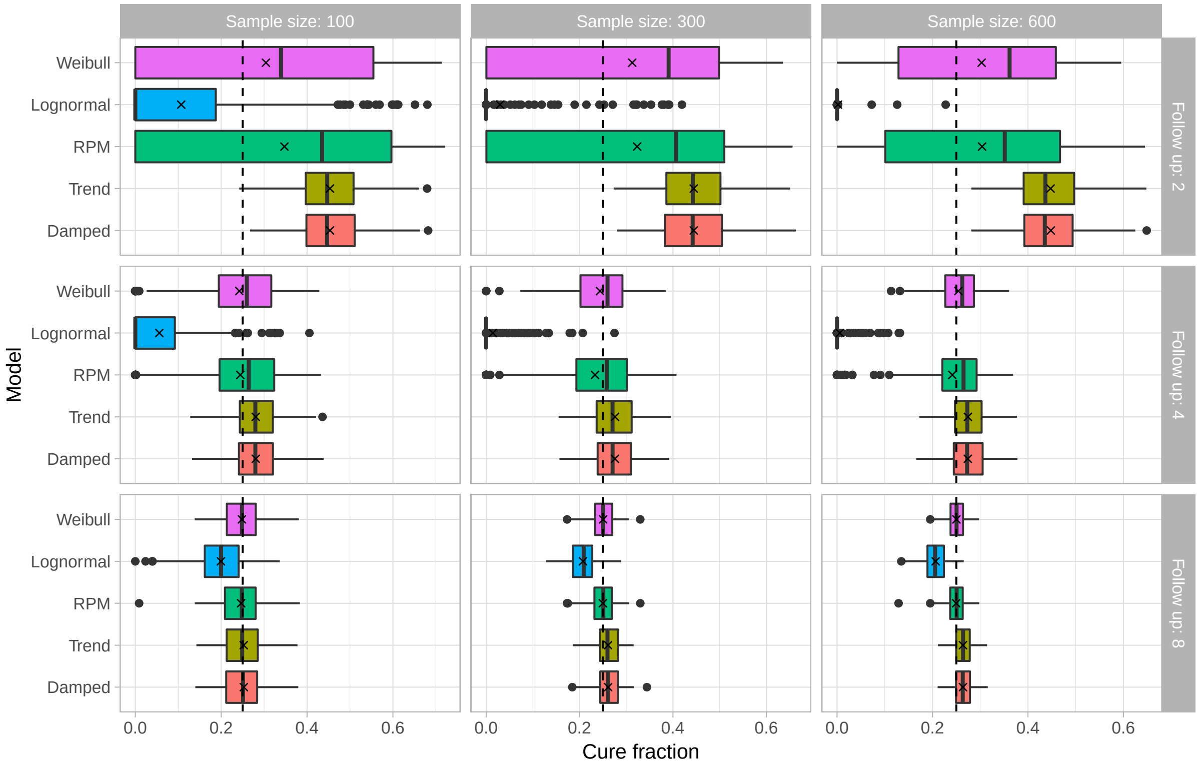
Estimates of the cure fraction for each model and each scenario are provided in Figure 5.5. Estimates are summarised using box and whisker plots. For these, the box spans the 25th and 75th percentiles, with whiskers extending to 1.5 times the inter-quartile range (range between 25th and 75th percentiles). Estimates outside the whiskers are shown as solid circles. Two summary values are displayed: the mean (cross) and median (line). The true value of 0.25 is denoted by a dashed line.

For the shortest follow-up none of the models provided accurate estimates of the cure fraction; the lognormal cure model underestimated the true value, with the remaining models overestimating the true value (based on their point estimates). As the length of follow-up increased estimates became more accurate: at the longest follow-up (when virtually all the uncured patients had died), the Weibull and RP cure models correctly identified the true cure fraction, with a very slight over-estimate for the two DCFMs (mean = 0.26 for both). In contrast, the lognormal cure model continued to under-estimate the cure fraction (mean = 0.20). In addition to increasing the accuracy of estimates, increasing follow-up also led to a decrease in variation. This was most notable for the Weibull and RP cure models. Increasing sample size also led to a decrease in variation, albeit to a lesser extent whilst also not affecting the accuracy of estimates.

In general, the Weibull and RP cure models provided slightly more accurate estimates of the true cure fraction than the two DCFMs. This did not however correspond to improved goodness of fit. This was most notable for the scenarios with shortest follow-up, for which both DCFMs had better overall MSE despite substantially over-estimating the cured fraction (range for mean estimates: DCFMs 44.5% to 45.3%, Weibull and RP cure models 30.3% to 34.7%).

For all nine scenarios the lognormal cure model tended to underestimate the cure fraction, often markedly. For example, with a sample size of 600 and two to four years follow-up, the mean cure fraction was 0.22% and 0.58% respectively. These estimates were between 30% to 45% (two years follow-up) and 24% to 27% (four years follow-up) smaller than estimates from the remaining models. One explanation for this underestimation may be because the non-cure fraction lognormal is able to adequately model the shape of the observed data. However, the more flexible DCFMs and RP cure model can also model the types of observed hazard patterns but had markedly different estimates of the cure fraction. Instead, this result appears to be because for short-to-medium follow-up the non-cure lognormal gives a slightly better fit to the observed data than a lognormal with a cure fraction. This is explored in more detail in Section 5.2.2

Figure 5.5: Estimates of the cure fraction



Dashed line shows true cure fraction (0.25). For each model, the cross shows the mean, the line shows the median

5.2.2 Non-cure models applied to data with a cure fraction

Visual extrapolations from the models which do not include a cured fraction (hence do not incorporate external information) are provided in Figure 5.6, along with DRSMs. The two DRSMs incorporate life tables but are misspecified as they do not allow for the possibility of a cured fraction. Due to the large number of models considered, this figure is split into two panes; for comparison current practice is included in both. Without external evidence none of the models provide accurate extrapolations. DRSMs include external evidence and avoid implausibly low extrapolations, but in general they also provide extrapolations that are more variable (with no notable improvement in accuracy) than their corresponding DSMs. The one exception is for the scenarios with the longest follow-up (eight years) for which the local trend DRSM provides accurate extrapolations. Estimates from the damped trend DRSM remain poor even at the longest follow-up. Neither current practice nor emerging practice models (without external information) were able to accurately predict the long-term hazard function for any of the scenarios considered. Estimates of bias and MSE are not quantified here as there is little merit in identifying the model with the best goodness of fit due to the generally poor extrapolation performance of the models considered. The poor performance of models without external evidence is not unexpected, as such models can only provide plausible extrapolations if the available data includes some of the long-term trend in the hazard. For this simulation study the long-term trend only becomes apparent towards the end of the longest follow-up when effectively all the uncured group is dead. In addition to providing less accurate extrapolations, models without external evidence also generally provided more variable extrapolations than models with external evidence.

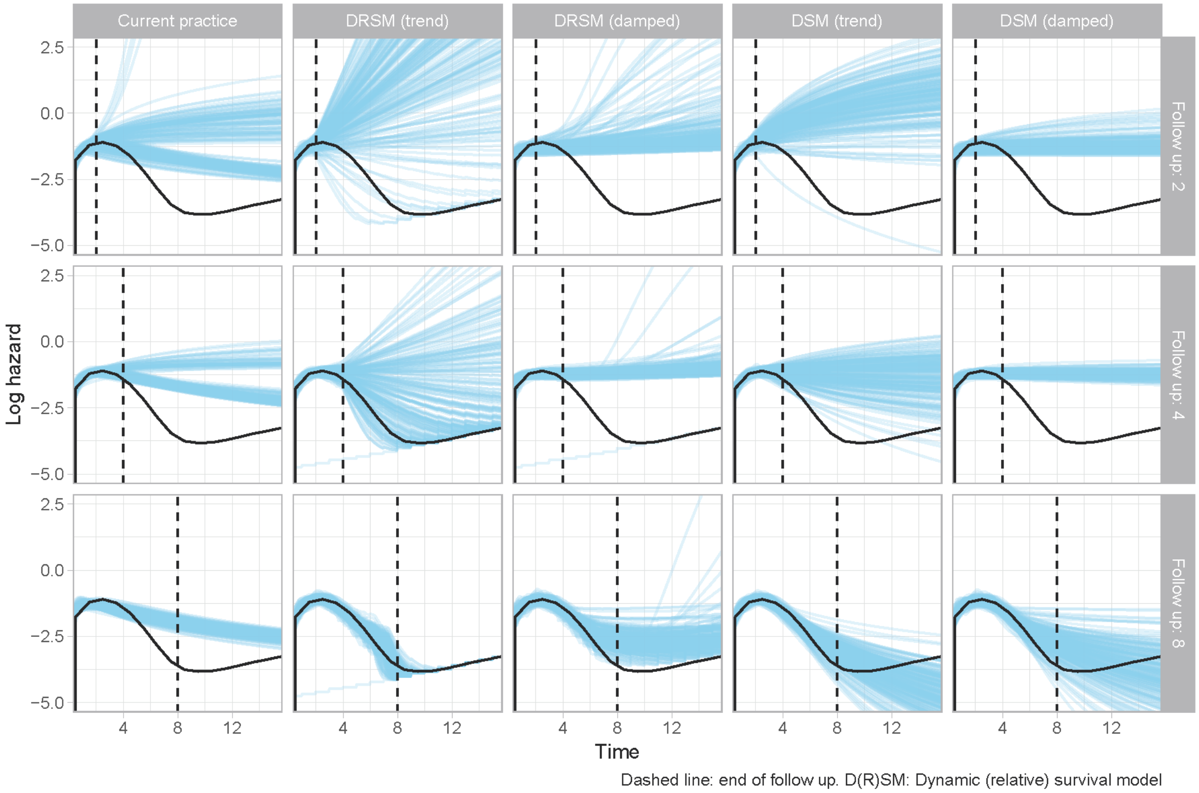
Qualitatively, the performance of models without external evidence was similar to the models in the previous chapter (which assessed the same models when the truth came from a mixture-Weibull model). For example, the flexible models provided better within-sample fit than the current practice models but higher variation in extrapolations. Again, the variation in extrapolations was largest for the GAM and FP(2) models, followed by the local trend model. In contrast to the previous chapter, the extrapolation performance of the FP(2) models was no worse than the GAMS (visually estimates are very similar), although this is primarily because in this simulation study extrapolations from both model classes lack face validity, with values much greater or smaller than the observed range of hazard values. In both simulation studies the damped trend DSM provided less variable extrapolations than the local trend model, the degree of dampening was the most pronounced in the six scenarios of this chapter with follow-up of two to four years.

Figure 5.6 shows three distinct extrapolation patterns from the current practice models. With the shortest follow-up some extrapolations increased very sharply towards infinity, whilst the majority of extrapolations demonstrated either a moderate increase or a moderate decrease.

With the middle follow-up the sharply increasing extrapolations vanished, whilst for the longest follow-up only the moderate decreasing extrapolations remained. These distinct patterns will arise from different parametric models; differences amongst the individual current practice models are explored in more depth in the next section.

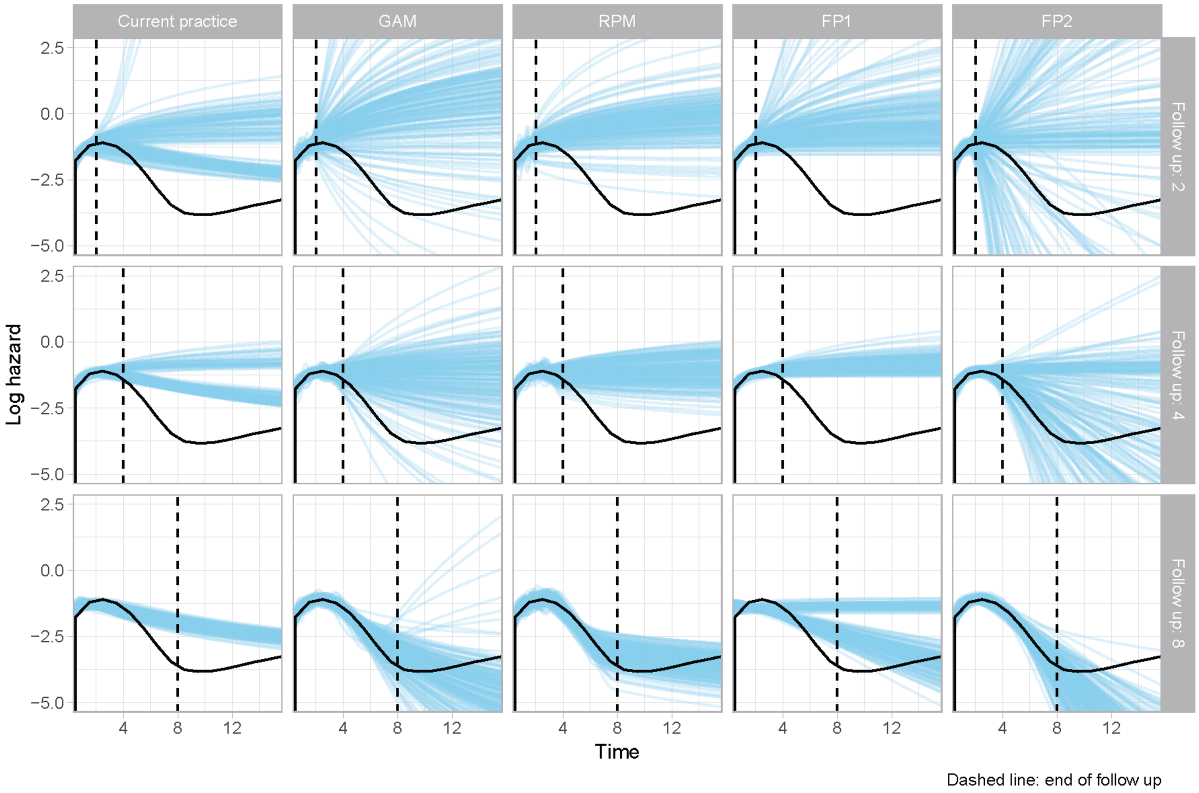
Figure 5.6: Estimates of the log-hazard compared to the truth: models without a cure fraction

(a) Current practice and dynamic models



Dashed line: end of follow up. D(R)SM: Dynamic (relative) survival model

(b) Current and emerging practice

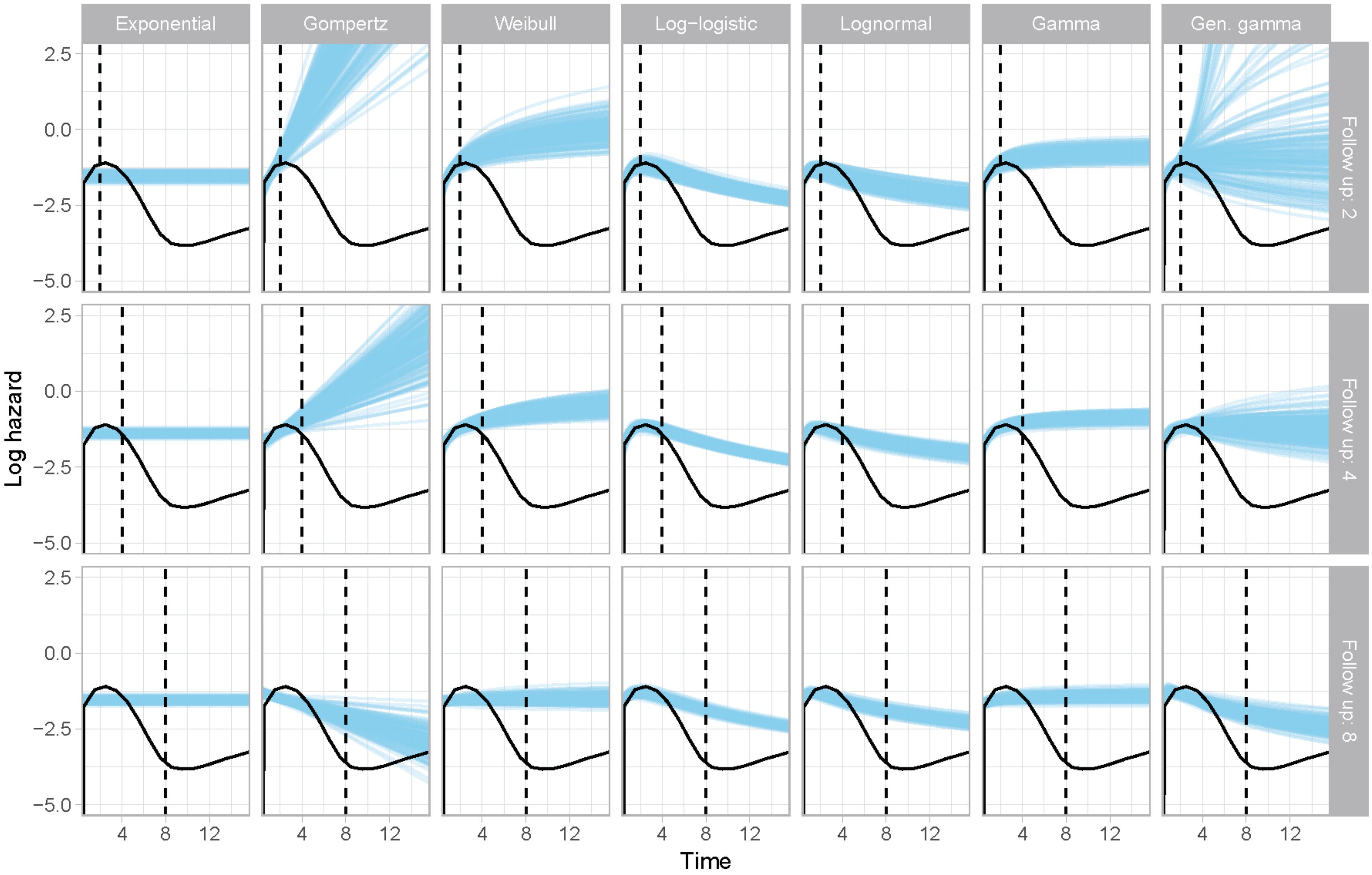


Dashed line: end of follow up

Non-cure current practice models

Visual estimates of both within-sample fit and extrapolations for the seven models comprising current practice are provided in Figure 5.7. For within-sample fit, all the models except for the exponential provide similar goodness of fit for follow-ups of two to four years (range in mean MSE values: 0.01 to 0.04), whilst for the longest follow-up none of the models provided good descriptions of the observed hazard function (range 0.28 for generalised gamma to 0.48 for gamma). For the scenarios with short-to-medium follow-up, despite all the models (excluding the exponential) having very similar within-sample fit, extrapolations varied markedly; extrapolations from the Gompertz, Weibull and gamma all increased, those from the log-logistic and lognormal decreased, whilst the generalised gamma provided extrapolations that could increase or decrease but on average were constant. This further highlights the potential futility in basing model choice on within-sample goodness of fit, as models with near-identical within-sample fit could provide qualitatively discrepant extrapolations (with no models providing accurate extrapolations). When comparing the Weibull and lognormal models with the corresponding cure fraction models, the cure models provided notably improved within-sample fit for the scenarios with the longest follow-up (range in improvement in MSE: 0.41 to 0.44 for the Weibull cure model and 0.14 to 0.18 for the lognormal cure model). For the scenarios with shorter follow-up there was very little difference in MSE values although the non-cure lognormal model typically provided lower bias values on average than the lognormal cure model. In practice MSE values are not available (these rely on knowing the truth), so model selection is based on goodness of fit measures such as AIC. The impact of choosing a cure or non-cure current practice specification on model selection is explored in the next section.

Figure 5.7: Estimates of the log-hazard compared to the truth: current practice (non-cure) models



Dashed line: end of follow up

Impact of omitting the cure fraction on model selection

As illustrated in Figure 5.6, incorrectly using a model without a cure fraction results in very poor visual extrapolations of the hazard function. Within-sample fit however appears generally acceptable, with the exception of the scenarios with the longest follow-up for the least complex models: current practice and FP(1). In practice there will be uncertainty about if the assumption of a cure fraction is tenable, so both models with and without a cure fraction may be considered. This subsection examines the impact of the choice of cure fraction on model selection. Here the focus is on AIC, as visual within-sample goodness of fit is often similar between cure and non-cure models, whilst extrapolations from many of the non-cure models may be deemed to be clinically plausible in the absence of a cure fraction. Table 5.3 provides the average absolute improvement in AIC by scenario when using a cure model in preference to its non-cure alternative. Dynamic models are omitted from Table 5.3 as their hierarchical structure means that there is no unique definition of the number of parameters (see Chapter 7 of Gelman *et al.* [177] for further details). On average, use of a cure model led to an improved within-sample fit for all nine scenarios, with larger improvements in more data-rich scenarios (longer follow-up and/or increased sample size). However, in the scenarios with the least data, there was very little difference between cure fraction models and their corresponding non-cure models, suggesting that it would be difficult to choose between the models in these situations. Given that extrapolations varied markedly between cure and non-cure models, this highlights the difficulties with model specification in data-poor scenarios.

Table 5.3: Improvement in AIC with cure model. Values are mean (95% confidence interval).

	Weibull		**Lognormal**		**Royston Parmar model**	
Two years						
n = 100	1.88	(0.88 to 5.68)	1.79	(0.74 to 6.57)	1.52	(0.13 to 3.69)
n = 300	8.17	(6.81 to 12.7)	9.01	(6.6 to 20.54)	7.8	(6.47 to 10.38)
n = 600	17.34	(15.89 to 22.35)	19.7	(15.71 to 30.43)	16.89	(15.64 to 19.59)
Four years						
n = 100	5.75	(2.91 to 12.37)	4.27	(2.58 to 11.16)	4.54	(2.77 to 7.8)
n = 300	18.35	(12.95 to 28.02)	16.74	(12.67 to 30.59)	15.33	(12.62 to 19.28)
n = 600	38.38	(29.87 to 54.7)	35.34	(28.06 to 52.82)	31.11	(27.75 to 34.65)
Eight years						
n = 100	29.08	(15.27 to 48.35)	13.32	(6.17 to 29.63)	10.66	(5.75 to 16.66)
n = 300	88.58	(67.86 to 114.82)	40.17	(27.73 to 60.16)	27.95	(21.74 to 34.17)
n = 600	176.12	(142.79 to 221)	80.28	(61.01 to 108.85)	53.26	(46.33 to 60.14)

5.2.3 Cure and relative survival models applied to data without a cure fraction

Models incorporating a cure fraction have been shown to provide a good fit to data with a cure fraction, which involve complex multi-model hazard shapes. A similar hazard pattern was used in Chapter 4, which was generated from a non-cure mixture of Weibull distributions. This section explores the fit of the cure models to the nine scenarios generated in Chapter 4, to compare their goodness of fit with the non-cure models. For completeness the two DRSMs are also considered.

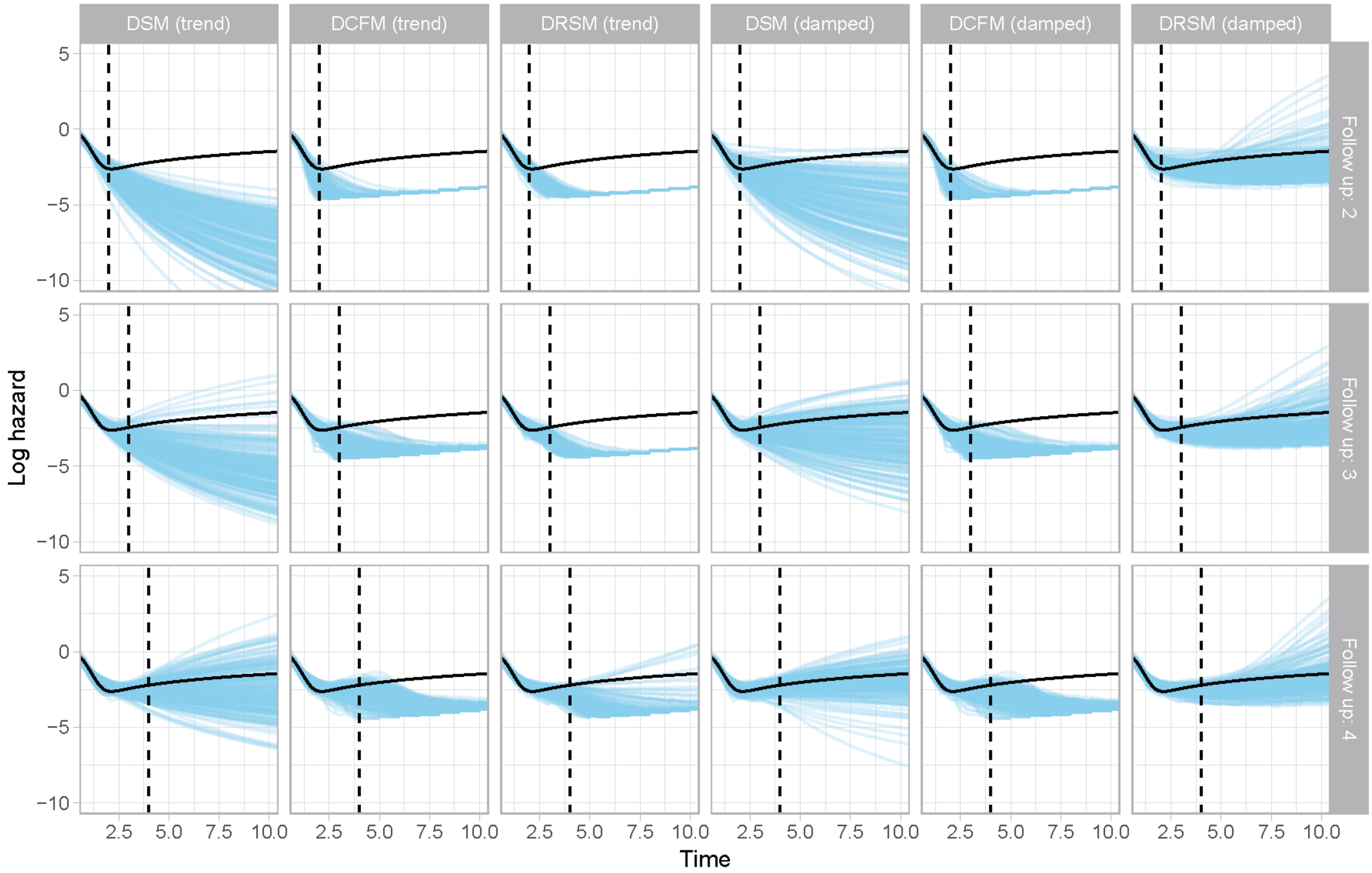
As with the other analyses in this and the previous chapter, varying length of follow-up had a larger impact than varying sample size. Visual goodness of fit for the within-sample and extrapolated phase are provided for the three lengths of follow-up with a sample size of 300 in Figure 5.8 for the dynamic models, and in Figure 5.9 for the current practice and Royston Parmar models. Both figures include models with a cure fraction and their corresponding models without external evidence (for comparison). The dynamic models also include the extension to relative survival models (DRSMs). Results for the remaining six scenarios (not displayed) were very similar to those presented here. Overall goodness of fit values, averaged across the within-sample and extrapolation periods, are provided in Table 5.4 for all nine scenarios. Within-sample fit (MSE and bias) was generally worst for the current practice and Weibull cure model, with the remaining models all providing similar within-sample estimates.

For the dynamic models, incorporating external evidence led to reduced variation in the extrapolations whilst also avoiding implausibly small extrapolations as they were constrained by the general population hazards. For data-poor scenarios these changes led to improved MSE values compared to dynamic models without external evidence. However, both DCFMs along with the local trend DRSM failed to identify the true long-term increase in the hazard function for any of the scenarios (with all three models providing very similar extrapolations). In the two scenarios with the richest data (follow-up of four years with a sample size of 300 or 600), the two dynamic models without external evidence had the lowest MSE values (of the six dynamic models). For the remaining seven scenarios, the damped trend DRSM had the lowest MSE values, with the second lowest MSE values for the local trend DRSM in five scenarios and the local trend DCFM in the remaining two. Results for bias were similar, with the damped trend DRSM being the least biased dynamic model in eight scenarios (and the second least biased, behind the damped trend DSM in the remaining scenario). The damped trend DRSM also had lower MSE and bias values than current practice models without external evidence in eight scenarios. For the remaining scenario, the difference in MSE (bias) values was 0.01 (0.06). In contrast, the remaining dynamic models with external evidence always had worse MSE and bias values than current practice models without external evidence.

None of the remaining three cure models (Weibull, lognormal, Royston-Parmar) were able to identify the long-term increase in the hazard for any of the nine scenarios. The overall goodness of fit of these models was similar to that of the DCFMs. The non-cure current practice model had a lower MSE and bias than the lognormal and Weibull cure models, although as discussed in the previous chapter this is primarily an artefact of the poor within-sample fit of the current practice non-cure model, combined with unrealistically low variation during the extrapolated phase. Differences between the two formulations of the RPM were most pronounced at the shortest follow-up (favouring the non-cure version), with little overall difference at follow-ups of three and four years.

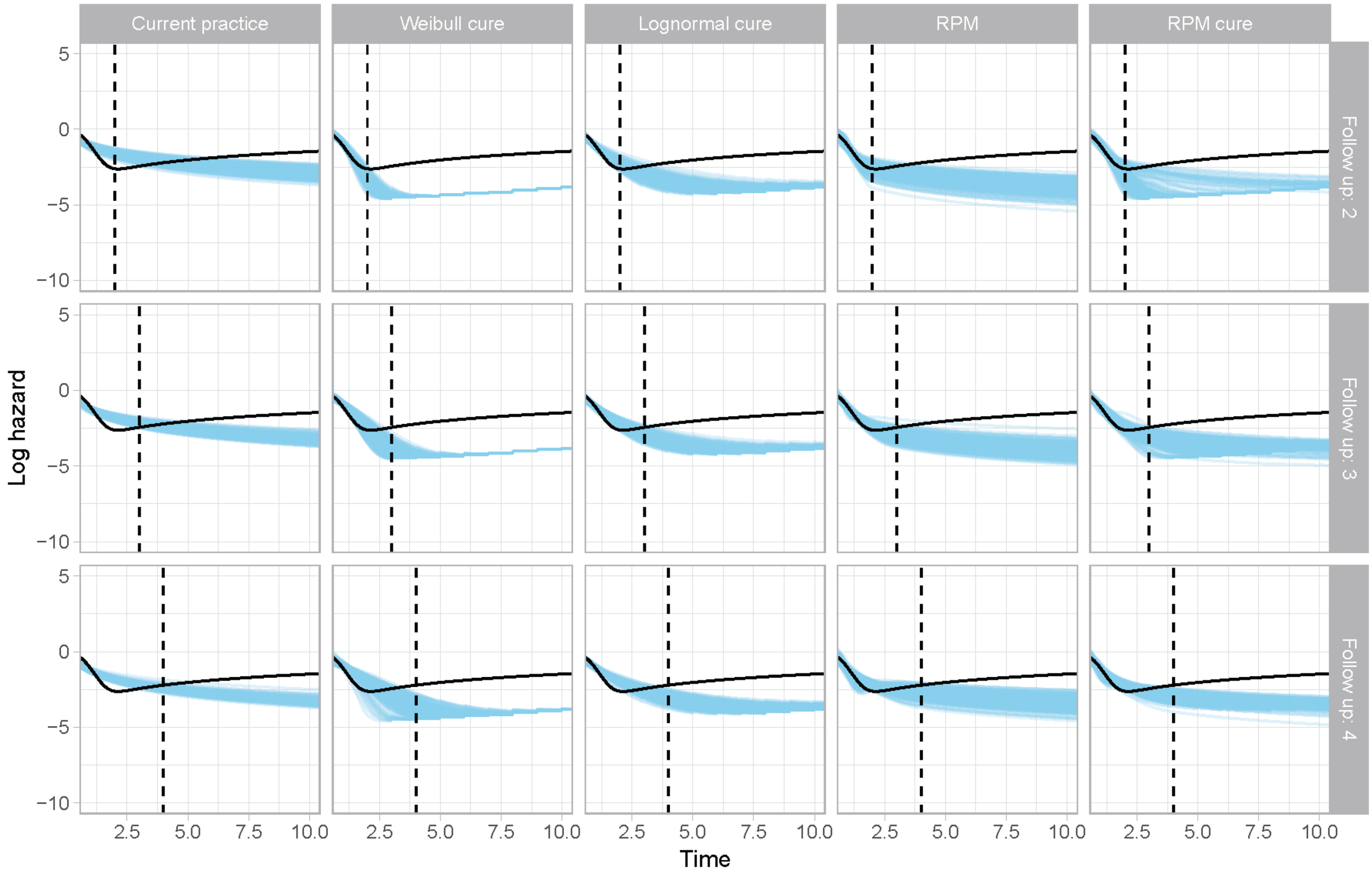
With the exception of RPM cure models, estimates of the cure fraction were similar for each of the cure models, with mean values ranging from 0.48 to 0.33 across the nine scenarios (for RPM cure models the range was 0.44 to 0.13). For all models increasing follow-up or sample size led to slightly lower estimates of the cure fraction.

Figure 5.8: Estimates of the log-hazard compared to the truth: dynamic models fit to non-cure data



Dashed line: end of follow up. DCFM: Dynamic cure fraction model. D(R)SM: Dynamic (relative) survival model.

Figure 5.9: Estimates of the log-hazard compared to the truth: current practice and Royston Parmar models fit to non-cure data



Dashed line: end of follow up

Table 5.4: Goodness of fit over the entire time horizon: cure and non-cure models fit to non-cure scenarios.

Overall mean squared error	**Sample size 100. Follow-up:**			**Sample size 300. Follow-up:**			**Sample size 600. Follow-up:**		
	2 years	**3 years**	**4 years**	**2 years**	**3 years**	**4 years**	**2 years**	**3 years**	**4 years**
Model	2.44	2.34	2.05	2.41	2.33	1.95	2.45	2.34	1.94
Weibull cure model	3.65	3.34	2.91	3.63	3.32	2.83	3.63	3.33	2.82
Royston-Parmar cure model	3.38	2.51	1.86	3.32	2.18	1.58	3.25	1.99	1.54
Local trend cure model	3.68	3.24	2.62	3.64	2.76	1.79	3.44	2.35	1.57
Damped trend cure model	3.68	3.24	2.61	3.64	2.76	1.78	3.45	2.35	1.54
Non cure models									
Current practice	1.01	1.19	1.26	0.94	1.15	1.19	0.90	1.12	1.15
Royston-Parmar model	1.98	2.38	1.87	2.21	2.38	1.50	2.25	2.36	1.40
Local trend, global level	6.03	7.12	4.27	15.61	6.67	1.36	18.04	4.65	0.57
Damped trend, global level	3.75	4.98	2.36	7.88	2.29	0.52	8.07	1.41	0.35
Overall bias									
Lognormal cure model	-1.15	-1.10	-0.96	-1.15	-1.11	-0.93	-1.17	-1.11	-0.93
Weibull cure model	-1.61	-1.44	-1.22	-1.61	-1.43	-1.19	-1.61	-1.44	-1.18
Royston-Parmar cure model	-1.51	-1.18	-0.94	-1.50	-1.09	-0.86	-1.48	-1.04	-0.86
Local trend cure model	-1.64	-1.47	-1.23	-1.64	-1.30	-0.90	-1.57	-1.16	-0.79
Damped trend cure model	-1.65	-1.47	-1.22	-1.64	-1.30	-0.90	-1.57	-1.16	-0.78
Non cure models									
Current practice	-0.36	-0.37	-0.35	-0.55	-0.55	-0.54	-0.60	-0.58	-0.56
Royston-Parmar model	-0.92	-1.07	-1.10	-1.07	-1.10	-1.11	-0.88	-0.79	-0.77
Local trend, global level	-1.36	-2.85	-3.13	-1.73	-1.72	-1.30	-1.31	-0.45	-0.11
Damped trend, global level	-0.35	-1.83	-1.80	-1.17	-0.78	-0.56	-0.76	-0.18	-0.14

5.3 Conclusions

The extrapolation performance of cure models, relative survival models (which incorporate external evidence) and non-cure models (which do not) was assessed in a simulation study where the truth included a cure fraction. The performance of the models incorporating external evidence was also assessed using the mixture Weibull simulation study of the previous chapter (which did not include a cure fraction).

For the cure fraction simulation study, use of models with a cure fraction led to generally acceptable extrapolations. In contrast, non-cure models failed to provide plausible extrapolations for up-to eight years of follow-up, when over three-quarters of the sample had died. Relative survival models improved on the performance of models without external evidence but did not perform as well as cure models. However, at the shortest follow-up considered (two years, with over a third of the sample dead), even the correctly specified Weibull cure model provided poor extrapolations (Figure 5.3). Of the cure models considered, the Weibull and lognormal models both make strong structural assumptions about the shape of the disease-specific hazard: that it is either monotonic or has a single turning point, respectively. The results shown here suggest that when these strong structural assumptions are incorrect, resulting extrapolations can be poor. In contrast, the cure RPM and DCFMs make very weak structural assumptions and provided extrapolations that were similar to the correctly specified model across the scenarios considered. In particular, the misspecified damped trend cure fraction model had the lowest overall MSE in four of the nine scenarios, and the second lowest in a further two. When applied to the mixture Weibull data, the models that incorporated external evidence provided extrapolations with reduced variability which avoided implausibly low estimates. However, most of these models failed to identify the true long-term increase in the hazard function. The one exception was the damped trend dynamic relative survival model, which often had the best overall goodness of fit of all the models considered.

For the cure fraction simulation study, the correctly specified Weibull cure model led to the most accurate estimates of the cure fraction for the cure models considered, on average. However, this did not result in improved goodness of fit (within-sample or extrapolations). Further, with short follow-up, estimates from the Weibull cure model were also highly variable whilst within-sample fit was no better than that from a non-cure Weibull model. Together these highlight the lack of identifiability for data with a cure fraction; where the cure fraction is unknown the overall observed hazard function may be described equally well by different combinations of disease-specific hazard functions and cure fractions.

There is limited assessment of the performance of cure fraction models in the literature. Using long-term registry data for ovarian cancer, Dickman and colleagues show that estimates of the cure fraction (their primary outcome of interest) was sensitive to model choice; varying from 0.059 for the lognormal cure model to 0.151 for the Weibull cure model [138]. The authors further note that, whilst the true cure fraction is unknown, the Weibull cure model provides more plausible estimates of the cure fraction. This is despite this model having worse within-sample fit (based on AIC), leading the authors to caution that "*The estimate of the cure fraction can be unstable when there are a small number at risk toward the end of follow-up*". The results of this chapter indicate that instability in estimates of the cure fraction are strongly influenced by the length of follow-up, with more variation when fewer deaths were observed. Grant and colleagues performed a simulation study to assess the performance of Weibull cure models when applied to data representative of a NICE appraisal [210]. Separate cure fractions were modelled for patients pre-progression and post-progression, with an overall cure fraction of 40% and the majority of uncured people dead by 150 months. Models fit to 40 months of follow-up were found to fit the observed data well but under-estimate the overall cure fraction and provide visually poor extrapolations. Hence the results of Grant and colleagues support the findings of this study in demonstrating that extrapolation with short follow-up (relative to lifetime follow-up) can provide poor extrapolations, even if within-sample fit is good

To this author's knowledge, this is the first time that the within-sample and extrapolation performance of different parametric cure models has been assessed. It is the first time that DCFMs have been considered, as this extension of DSMs is a novel development of this thesis. Another strength of this study is the assessment of the impact that model misspecification has on both goodness of fit and model selection in practice. This includes both misspecifying the model for disease-specific mortality in a cure fraction model, and the misspecification of not using a cure fraction model for data with a cure fraction. For the former, use of a misspecified current practice model did not affect within-sample fit but led to very poor and consistently biased extrapolations. In contrast, use of a misspecified DCFM had little effect. For all model classes, use of a model without external evidence provided extremely poor extrapolations.

A potential limitation of this study is that the results realistically only represent an upper-bound on the performance of the cure models in practice. This is because for this study it has been assumed that the survival of the cured patients is known with certainty (the same life tables are used in the data-generating mechanism and the models). In reality, there is likely to be some misspecification in this; individual patient characteristics and local geographical factors may lead to survival that is different to national life tables. This would affect absolute goodness of fit but is unlikely to affect the relative performance of the models assessed. The hazard function of the uncured population is also relatively simple, arising from a monotonic Weibull model. In reality the hazard function may not be monotonic which again may hamper model performance. This is likely to most affect the Weibull and lognormal cure models, as this study has demonstrated that these extensions to current practice models are sensitive to model misspecification. This simulation study also only assessed one instance of one data-generating mechanism. Future research could continue to assess the goodness of fit of cure models with different data-generating mechanisms, or in situations with real data with long follow-up where a proportion of the sample are known to be cured. This could include situations where the 'cured' fraction have a mortality that is persistently elevated compared to the general population. This elevated risk could be due to prior treatment effects, or if there are underlying risk factors for other-cause mortality that are not addressed by the curative treatment. In addition, it would be beneficial to know if cure versions of RPMs are affected by misspecification (do not include the true model as a special case). Given the promising performance of GAMs in the mixture-Weibull (non-cure) simulation study, extensions to this model to incorporate a cure fraction would also be of interest. In the previous simulation study GAMs provided more variable extrapolations than dynamic models, so it would be interesting to see how the inclusion of external data affected this variation. Future studies could also expand the data-generating mechanism to consider the impact of disease progression on both survival and censoring, as was modelled in the study of Grant and colleagues [210].

The presence of a cure fraction creates complex hazard patterns which can pose a challenge for extrapolation. Extensions of current practice models to incorporate a cure fraction work well if they match the true data-generating mechanism but can provide poor results otherwise. Dynamic models with a cure fraction generally performed as well as the correctly specified model, whilst avoiding the sensitivity to model misspecification. For all the models evaluated, incorrectly omitting the cure fraction led to very poor extrapolations, whilst the usefulness of cure models was limited when the true data-generating mechanism did not involve a cure fraction. When the truth did include a cure fraction, all the cure models provided poor extrapolations at the shortest follow-up considered, and in the data-poor scenarios there was little difference in the within-sample fit of cure and non-cure models. Hence incorporating external data, in the form of general population hazards, can in some situations improve extrapolation performance but it is not guaranteed to do so, and it is not a substitute for having adequate follow-up. A limitation of this and the previous chapter is that neither considered the use of dynamic models in the context of a technology appraisal. This evidence gap is addressed in the next chapter.

Chapter 6

Dynamic survival models in health technology assessment: a case study

Survival models, including DSMs, are used to generate estimates of transition probabilities for use in HTA. The previous two chapters have examined the extrapolation performance of DSMs and compared this to alternative survival models. In contrast, this chapter demonstrates the use of DSMs in the context of HTA, when extrapolations impact upon estimates of overall costs and QALYs and hence cost-effectiveness. To demonstrate this, this chapter presents a re-analysis of an existing NICE appraisal focusing on the extrapolation of overall survival (OS) data. In the original appraisal current practice models and RPMs were both considered. As shall be seen, estimates of cost-effectiveness were sensitive to the choice of extrapolating model, and there was disagreement between the approach favoured by the submitting company and the approach of the evidence review group (ERG), who provided an independent critique of the company submission. For this case-study, extrapolations are re-estimated using DSMs; it is contended that a key strength of the dynamic models in these situations is the ability to base model choice on the anticipated clinical characteristics of survival (or mortality) during the extrapolated phase. In addition, a key criticism of the company's original approach to extrapolation was that eventually the extrapolated hazards fell below that of the age-matched general population. This motivates the use of DRSMs, first introduced in Section 3.3, which incorporate general population hazards as external data and ensure that extrapolations never fall below these values.

This chapter begins with an overview of the pivotal trial used for the extrapolation of OS data, along with the extrapolation approaches employed by the submitting company and the ERG. Extrapolations arising from DSMs and DRSMs are then presented, followed by a replication of the original health economic model. The impact of dynamic models on estimates of cost-effectiveness and the uncertainty in these estimates are then discussed, as well as an extension to incorporate time-varying treatment effects.

6.1 Case study: squamous non-small-cell lung cancer

The existing HTA was a submission to NICE as part of their TA programme [213]. As part of these appraisals, a company submits evidence on the clinical- and cost-effectiveness of a technology. This evidence is subject to a critique by an independent ERG. A NICE committee considers both the company submission and ERG critique as part of their decision-making process. The NICE committee provides recommendations on whether or not the technology is judged to be cost-effective and produces a recommendation of whether the technology should be recommended for routine use.

For this appraisal, the population of interest was people with previously treated locally advanced or metastatic (stage IIIB or IV) squamous non-small-cell lung cancer. The intervention was nivolumab and the sole comparator in the company's submission was docetaxel. The main evidence source was the pivotal open-label phase III trial CheckMate-017 (NCT01642004) which compared nivolumab (n = 135) against docetaxel (n = 137) for the population of interest (whose previous treatment was with platinum combination chemotherapy) [214]. Patient follow-up was between 11 and 24 months. At the end of follow-up there had been 86 (63.7%) and 113 (82.5%) deaths in the nivolumab and docetaxel arms, respectively. Treatment was until either disease progression or unacceptable levels of toxicity. The primary outcome measure was OS. Evidence on effectiveness come solely from this trial and there was no treatment switching in the data used in the original company's submission.

6.1.1 Company's approach to extrapolation

Extrapolations were obtained for both OS and progression-free survival (PFS). For both outcomes, the company based their approach to extrapolation on the guidance in NICE TSD 14 [20]. First, the assumption of proportional hazards was checked both visually and using significance testing. The company considered both current practice survival models and RPMs. For the former, it is unclear if the company considered all seven models described in TSD 14 (exponential, Weibull, Gompertz, log-logistic, lognormal, gamma, and generalised gamma), as results were only shown for the log-logistic, lognormal and generalised gamma models. RPMs with up to two internal knots modelled on the hazard, normal and odds scales (corresponding to extensions of the Weibull, lognormal and log-logistic models, respectively) were considered. Both AIC and BIC were considered for goodness of fit. The plausibility of extrapolations was assessed against external evidence.

For OS, the assumption of proportional hazards appeared to hold. For docetaxel, the log-logistic model was deemed to be the best with regards to both within-sample goodness of fit and plausibility. The treatment effect for nivolumab was modelled as a hazard ratio. It is noted that treatment effects within a log-logistic model do not have a proportional hazards interpretation. An alternative approach to analysing the treatment effect is provided in Section 6.4.1. For PFS, the proportional hazards assumption was judged to be violated. Hence the company modelled both treatments using an RPM with two internal knots on the hazard scale.

The probabilistic base-case incremental cost-effectiveness ratio (ICER) arising from this approach was £86,000 (for this chapter, all ICERs discussed in text are given to the nearest £500 and are per QALY gained), with a survival gain of 1.31 years for nivolumab [213]. This value was robust to alternative approaches to extrapolation for PFS, but not for OS. For example, when varying the hazard ratio across its plausible range the ICER varied from £55,000 to £169,000.

6.1.2 Independent critique

The independent ERG were critical of the company's OS extrapolations, in particular the fact that the extrapolated hazard (mortality rate) eventually fell below that of the age-matched general population:

> *"The ERG considers this to be wholly implausible, and inconsistent with any clinical evidence of treating metastatic disease"* [215] p88.

The ERG contented that the extrapolated hazard for OS was likely to increase over time due to ageing. Despite this, they extrapolated a constant hazard over time (using an exponential model). This was fit from 40 weeks (9.2 months) of follow-up, with the ERG suggesting that this cut-off was supported by the data. The ERG's approach to OS extrapolation increased the company's base-case ICER from £86,000 to £132,000, whilst the estimated lifetime survival gain more than halved, from 1.31 to 0.64 years. In response, the company amended their extrapolation approach to cap the extrapolated hazard rate so that it never fell below that of the corresponding general population mortality. The company's revised base-case ICER was £92,000, with a survival benefit of 1.16 years [216]. However, the ERG were critical of the company's revised approach:

> *"This change does not address the fundamental problem, that the log-logistic curve can never accurately represent a human population which in the long-term always experiences steadily increasing mortality hazards with advancing age"* [217] p2.

To summarise, the approach to extrapolating OS was identified as both a key area of uncertainty and a key driver of estimates of cost-effectiveness. The company fit survival models to all the available data and extrapolated a decreasing trend in the hazard. In contrast, the ERG fit a survival model to a subset of the available data and extrapolated a constant value (no trend), whilst also criticising the company's original extrapolations for eventually falling below that of the age-sex matched general population. The company in turn criticised the ERG's approach as ignoring the trend in the hazard observed in the trial and lacking robustness by not using all the available data.

6.2 Re-analysis of the clinical effectiveness data

6.2.1 Analysis without external evidence

Data on OS were digitised from the pivotal trial publication [214] using Engauge digitiser [218]. These digitised data were used to replicate the original individual patient data using the algorithm of Guyot and colleagues, as implemented in the R package `survHE` [219, 220].

For consistency with the original company submission, the treatment effect for nivolumab is modelled as a hazard ratio compared with docetaxel, for which both current practice and RP models are considered. For RPMs, up to five internal knots are considered. Within-sample goodness of fit is measured using AIC (there were no substantial differences when using BIC instead of AIC). As with the previous chapter, values of the IER (relative goodness of fit, which is 100% for the best-fitting model) are also provided. Results are provided in Tables 6.1 and 6.2, arranged by AIC.

Table 6.1: Goodness of fit statistics for the current practice models.

Model	**Log-logistic**	**Lognormal**	**Generalized gamma**	**Gamma**	**Weibull**	**Exponential**	**Gompertz**
AIC	721.3	723.0	723.6	724.9	726.9	730.8	732.0
IER	100%	42%	31%	17%	6%	1%	0%

Table 6.2: Goodness of fit statistics for the Royston Parmar models.

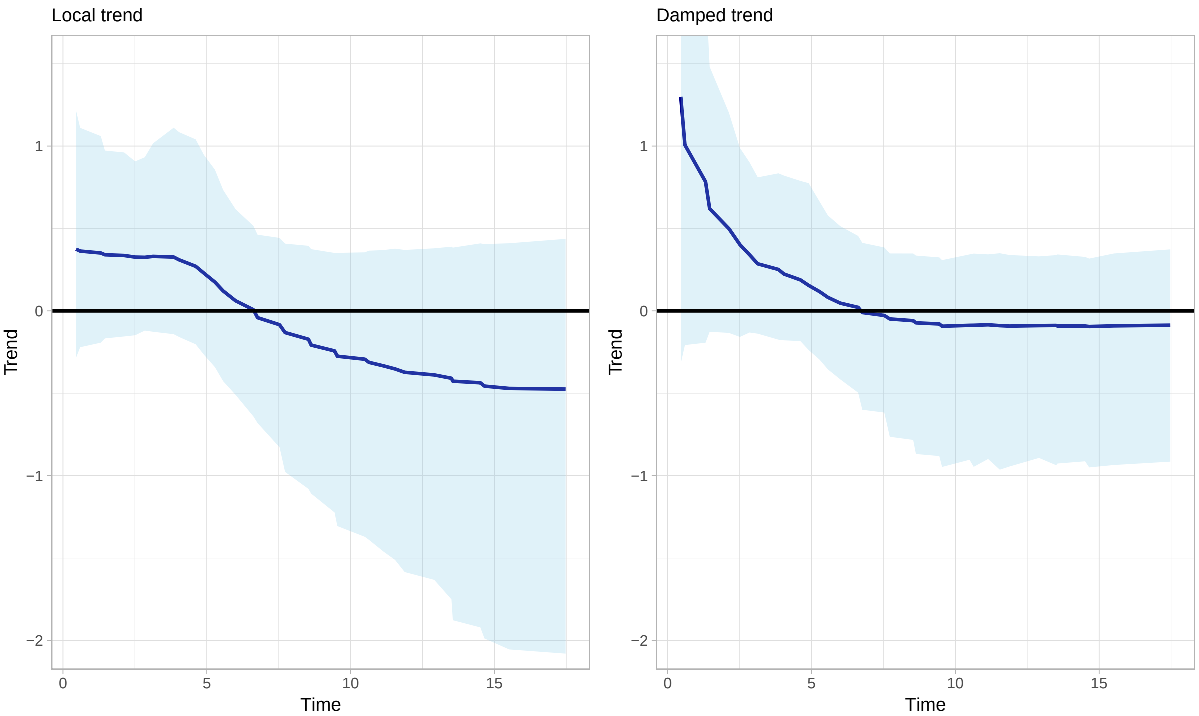
Interior knots	**AIC: Hazard**	**AIC: Normal**	**AIC: Odds**	**IER: Hazard**	**IER: Normal**	**IER: Odds**
None	726.9	723.0	721.3	6%	42%	100%
One	724.5	722.8	722.9	19%	45%	44%
Two	724.4	724.0	724.8	21%	25%	17%
Three	724.1	722.6	723.1	24%	50%	39%
Four	722.4	722.7	722.4	55%	48%	57%
Five	725.8	725.9	725.7	11%	10%	11%

The results show that the log-logistic model is both the best-fitting (based on AIC) of the standard survival models and of the more flexible RPMs (as an RPM on the odds scale with no interior knots is equivalent to a log-logistic model). Estimates of the hazard function from the second-best fitting RPM (four internal knots, odds scale) were visually very similar to the log-logistic model for both the within-sample and extrapolated periods (estimates not shown).

Two DSMs are evaluated: a local trend and a damped trend model, both with a global level. The focus of this case-study is on comparing dynamic models with the models used in the original appraisal (current practice models and RPMs). An advantage of using a DSM is that it allows for a direct estimate of the trend in the hazard function over time, along with the uncertainty in this estimate. This is of particular use in this case-study, as there was disagreement on the observed trend at the end of follow-up, with the company modelling a decreasing trend and the ERG modelling no trend in the hazard (by assuming a constant hazard). Figure 6.1 displays the time-varying estimate of the trend from the two DSMs. The trend estimate from both models is initially positive followed by a decrease. For the local trend model this decrease starts at about four months, whilst for the damped trend model this decrease starts immediately. For both models the trend becomes negative at about half a year. For the local trend model the trend estimates continue to decrease, albeit with a large degree of uncertainty. For the damped trend model the trend is almost zero after half a year, suggesting that after this time the assumption of a constant hazard may be appropriate. Neither of the dynamic models estimated a monotonic trend, suggesting that models which assume monotonicity (such as the Weibull and Gompertz) are inappropriate. In contrast, use of a log-logistic or lognormal model may be acceptable, as the hazards from these can increase then decrease. Further, the confidence intervals from both models include zero at all time points, indicating that a constant hazard model cannot be ruled out.

A visual comparison of the fit from the two DSMs along with the original company approach (log-logistic) and ERG approach (hybrid exponential) is provided in Figure 6.2. For the extrapolations, estimates of the annual hazard of all-cause mortality for the age-matched general population are also included. These estimates came from lifetables supplied by the Human Mortality Database [200] and are 2016 data for the UK, assuming a starting age of 63 (the median

Figure 6.1: Estimates of the trend in the hazard function from two dynamic survival models.

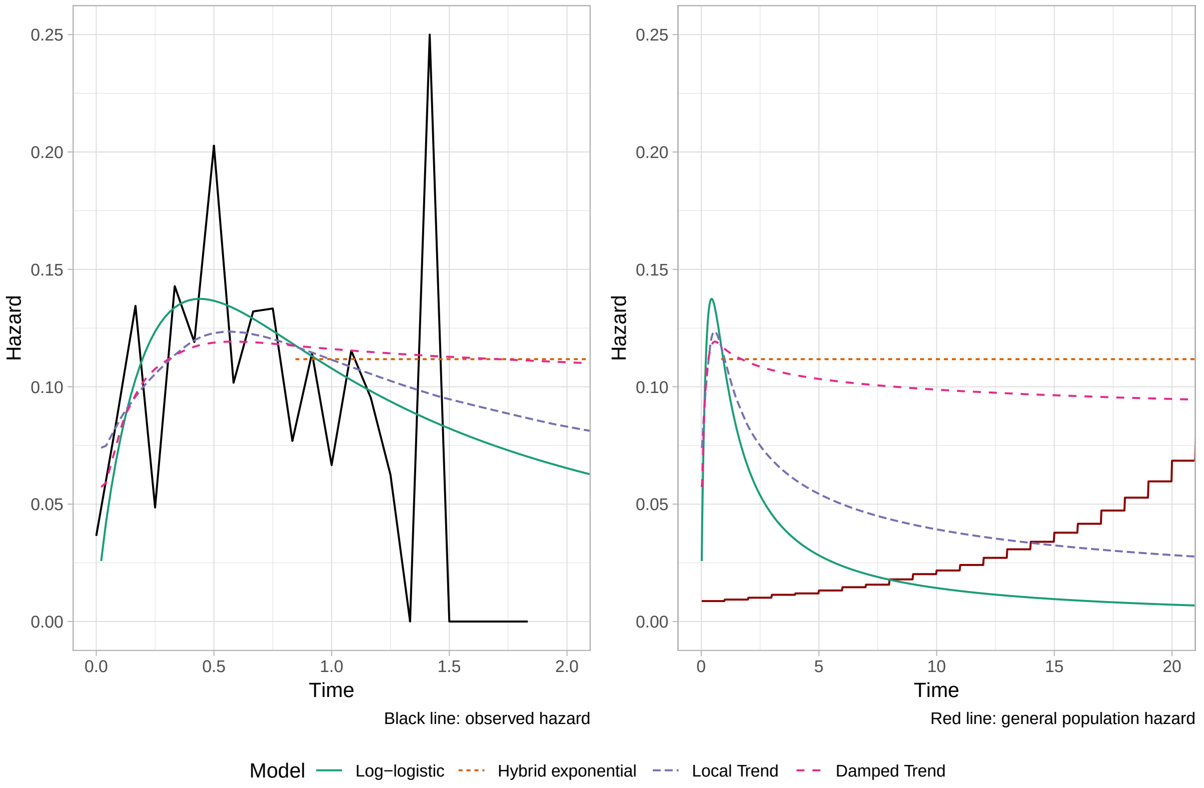


Solid-blue line: point-estimates, with 95% confidence intervals in pale blue. Black line = no trend.

age of participants in the pivotal CheckMate 017 trial). For the first year of follow-up estimates of the hazard function from the log-logistic and two dynamic models are visually similar, albeit the peak in the hazard (which occurs between four and six months for these three models) is more pronounced for the log-logistic. At one year of follow-up there are only 30 people still at risk (22% of the starting sample); this small sample size may be driving the difference in estimates from the three models observed after one-year. These differences continue into the extrapolated phase; the current practice log-logistic model estimates the strongest decrease in the hazard function. In contrast, the damped trend model estimates almost constant hazards; in the short-term these estimates are very similar to those from the ERG approach, but they become increasingly smaller than the ERG extrapolations as the time horizon increases. Two advantages of using the dynamic model in preference to the ERG approach are that first extrapolations are based on all the data (using the ERG approach only a third of the original sample, 45 people, contribute to hazard estimates), and secondly the DSM avoids the arbitrary choice of a cut-point.

Extrapolations from the local trend model fall between the log-logistic and damped trend models, eventually falling below corresponding age-matched estimates from the general population at approximately 15 years. Hence, long-term extrapolations from the local trend model are implausible for the outcome of all-cause mortality. It is possible to provide a post-hoc correction to the extrapolations to ensure that they never fall below general population hazards, but as noted by the ERG this approach is not entirely satisfactory. A better approach would be to formally incorporate this relevant external evidence within the statistical models to ensure that extrapolations are plausible. These can be achieved using the framework of relative survival models (Section 6.2.2).

Figure 6.2: Hazard estimates. Left: within-sample, right: extrapolations.



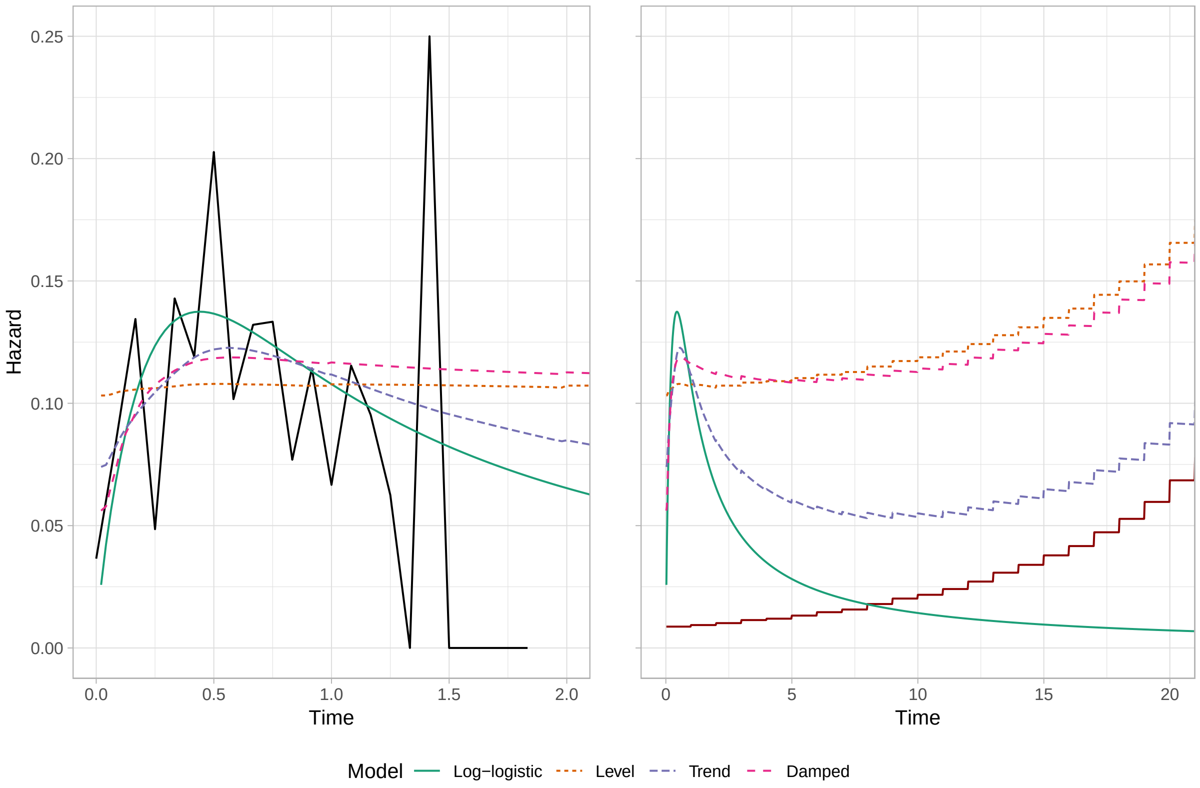
Black line: observed hazard. Red line: general population hazard.

6.2.2 Analysis with external evidence

External evidence, in the form of general population lifetables, are included in the dynamic models via DRSMs. The technical specification for these is provided in Section 3.3. An alternative approach would be to incorporate these within dynamic cure-fraction models as described in the same section and implemented in Section 5. Cure models were not considered as clinical input (in the original submission) did not mention cure as a possibility. In contrast, DRSMs will ensure that extrapolated hazards never fall below that of the general population hazards. This makes their use ideal in this case-study where this aspect of long-term plausibility had previously been highlighted.

Three DRSMs were evaluated: local trend, damped trend, and local level implementations. Estimates from these are provided in Figure 6.3. Within-sample estimates from the two trend models are very similar to the corresponding dynamic models which do not include external evidence. Visually, the local level DRSM provides within-sample estimates that are similar to an exponential model and do not fit the data as well as the other DRSMs. This suggests that, without a trend, the local level model would require large variation in the level to closely match the observed data. If estimates from the local level DRSM were more variable then there is a danger that they could over-fit the data, particularly as follow-up time increases and the sample size decreases. Extrapolations from the local level and damped trend DRSMs are very similar to each other, showing that (as with the damped trend DSM) there is a pronounced dampening of the trend before the end of follow-up. After 20 years, hazards from all the DRSMs are greater than the general population estimates, implying that there is a non-negligible extrapolated excess hazard. After about ten years the local trend DRSM extrapolates an increasing hazard, suggesting that after this point the influence of ageing on the hazard function outweighs the extrapolated decrease in the excess hazard.

Figure 6.3: Hazard estimates. Left: within-sample, right: extrapolations.



Black line: observed hazard. Red line: general population hazard.

6.3 Re-analysis of the cost-effectiveness data

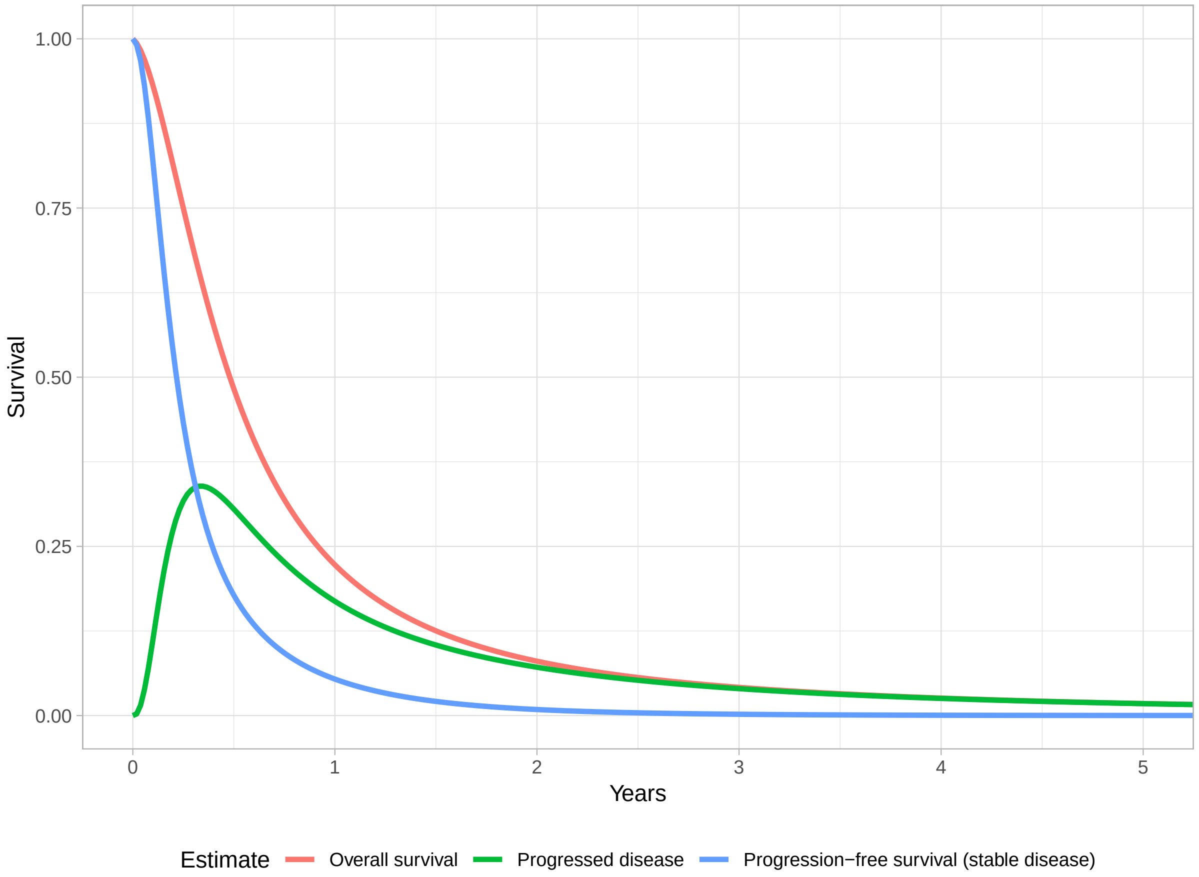
6.3.1 Health economic model

The company submitted a three-state partitioned survival economic model, with the states 'progression-free', 'progressed disease', and 'death'. State membership can be derived from lifetime estimates of OS and PFS:

- The area under the PFS curve corresponds to progression-free state membership.
- The difference between the area under the OS curve and the area under the PFS curve corresponds to progressed disease state membership.
- The area above the overall survival curve corresponds to membership of the death health state.

An example of the survival estimates is provided in Figure 6.4. The time-horizon for the economic evaluation was 20 years, assumed to represent a lifetime (the median age of trial participants was 63 years), with a 1-week time cycle. Utility data and resource use were primarily taken from CheckMate-017 [213].

Figure 6.4: Example of state membership in a partitioned survival model.



The health economic model uses discrete time cycles. Let D_t, S_t and P_t be the proportion of patients in the dead, stable, and progressed health states at time t, respectively. Hence, the proportion of the cohort in each health state over time was calculated as follows:

- Hazard estimates (λ) for OS and PFS were converted to transition probabilities using the formula: $p = 1 - \exp(-\lambda \times \tau)$, where τ is the cycle length.
- $D_t = D_{t-1} + (S_{t-1} + P_{t-1}) \times p_t^{\text{OS}}$
- $S_t = \min(S_{t-1} \times (1 - p_t^{\text{PFS}}), 1 - D_t)$

- $P_t = 1 - (D_t + S_t)$

For this re-analysis a partitioned survival analysis model was used for consistency with the original company submission. An advantage of this economic modelling approach is that it can be used with digitised patient level data. However, it should be noted that there are a number of limitations associated with this approach [221]. This includes two simplifying assumptions when implementing a partitioned survival analysis model. First that the probability of death is independent of a patient's health state, secondly that the OS and PFS curves are independent. Neither assumption is likely to be realistic. These limitations are inherent to partitioned survival analysis models and are independent of the statistical approach used for extrapolation. As such, they are noted as limitations, but not addressed here as this is beyond the scope of this thesis.

An overview of the costs included are provided in Table 6.3. Implementation of the PSA was based on the company approach of using a Gamma distribution, assuming that the standard error was 10% of the mean. In this replication, drug acquisition costs are only applied for the stable disease health state because evidence for progressed disease were submitted as confidential-in-confidence and hence were not publicly available.

Table 6.3: Costs included in the health economic model.

Description	**Mean value (£)**
Stable disease costs	
Disease management (per 4 weeks)	313.55
Drug acquisition: nivolumab (per 2 weeks)	2,634
Drug acquisition: docetaxel (per 3 weeks)	900
Administration: nivolumab	269.92
Administration: docetaxel	167.34
Monitoring (per 4 weeks)	151.89
Progressed disease	
Disease management (per 4 weeks)	766.62
Drug acquisition cost	Not reported
Transition to death	
End of life care	3,628.70

Data on the frequency of adverse events were also marked as commercial-in-confidence. A breakdown of results from the company's submission states that, of the average discounted costs per patient, adverse events contributed £228 and £1,304 for nivolumab and docetaxel, respectively [213]. As such, the average discounted results from the replicated model were adjusted by these amounts to reflect the impact of adverse events. As with costs, a post-hoc adjustment was made to the model outputs from the replicated model to reflect the impact of adverse events. These were to decrease the average QALYs by 0.01 and 0.05 for nivolumab and docetaxel, respectively.

The utility values attributed to the two alive health states were 0.750 and 0.592 for the stable and progressed states, with standard deviations of 0.236 and 0.315, respectively. Details on the distribution used in company's PSA were not provided. Hence, the available data (mean and standard deviation) were used to derive parameters of the beta distribution using the method of moments [222].

6.3.2 Replicating the original submission

The company's model was replicated in R, a comparison of the cost-effectiveness results is provided in Table 6.4. The comparison is against the original company submission (using their approach to extrapolation) as this had the most detailed reporting of results. For OS this approach used a log-logistic for docetaxel and applied a hazard ratio for nivolumab. For PFS RPMs were used.

Given that the individual patient level data were not available and these data were recreated, identical results are not expected. There is in general close agreement between the original company submission and the replication for the deterministic results. There is a slight under-estimation of the mean life-years with nivolumab (2.22 vs 2.26) and hence QALYs (1.28 vs 1.30). The corresponding values for docetaxel were identical to two decimal places when comparing the original and replicated submission. For both treatments there was also a slight under-estimation of absolute costs (by about £500 and £1,000 for nivolumab and docetaxel, respectively) however, it is known that the replication will slightly under-estimate absolute costs, as it was not possible to include a drug acquisition cost for the progressed disease health state. Patients in the docetaxel arm spend about twice as long in the progressed health state (after time is discounted), which matches with the under-estimate being about twice as large for this health state than for nivolumab. The probabilistic results values are also similar to the deterministic results. The number of probabilistic iterations (samples) used in the company submission was 1,000. For the replication it was 2,000 to ensure that results were stable.

Table 6.4: Comparison of cost-effectiveness results: original submission and replication.

	Absolute Value			**Incremental values**		**ICER**
Deterministic results	**Life Years**	**QALYs**	**Cost**	**QALYs**	**Cost**	
Original submission						
Nivolumab	2.26	1.3	£86,599	0.76	£65,355	£85,950
Docetaxel	0.95	0.54	£21,243			
De novo replication						
Nivolumab	2.22	1.28	£86,073	0.74	£65,891	£89,309
Docetaxel	0.95	0.54	£20,182			
Probabilistic results	**Life Years**	**QALYs**	**Cost**	**QALYs**	**Cost**	**ICER**
Original submission						
Nivolumab	NR	1.35	£91,677	0.77	£68,938	£89,343
Docetaxel	NR	0.58	£22,739			
De novo replication						
Nivolumab	-	1.29	£85,882	0.74	£65,470	£87,926
Docetaxel	-	0.55	£20,413			

ICER = incremental cost-effectiveness ratio = incremental costs / incremental QALYs. All values are discounted with the exception of life-years. NR: Not reported.

6.3.3 Results for dynamic survival models

Results are provided in Table 6.5 for two DSMs (which do not incorporate data on general population mortality) and three DRSMs (which do). For comparison, three replicated analyses are also shown. These are:

- The company's original submission (extrapolation with a log-logistic model)
- The company's original submission, with extrapolations capped so that hazard estimates never fall below those of the general population.
- The ERG's approach (use Kaplan-Meier estimates up to 40 weeks, fit an exponential to the remaining data and use to extrapolate).

As shown in Table 6.5 and in Figures 6.2 and 6.3, extrapolations can differ between the five dynamic models; this in turn leads to variation in the cost-effectiveness results. Of the dynamic models, the largest predicted number of QALYs gained for the nivolumab cohort occurs for the local trend DSM (1.06), but as previously noted the extrapolated hazards from this model eventually fall below that of the age-matched general population, which is likely to be clinically implausible. The smallest numbers of QALYs gained for the nivolumab cohort occurs for both damped trend dynamic models and the local level DRSM (0.86 to 0.88). These three models all extrapolate a near-constant hazard (or excess hazard). Of the dynamic models, the three incorporating external evidence (general population mortality) are likely to provide the most plausible extrapolations. Incremental QALYs arising from the three DRSMs range from 0.34 (damped trend) to 0.45 (local trend). This reflects the different extrapolations from these models (see Figure 6.3) and is reflected in ICERs ranging from £142,825 to £122,328 respectively. Whilst estimates of the incremental QALY gain varied notably between the DRSMs, there was a much smaller variation in estimates of the incremental cost. This is because the majority of the costs are incurred during the start of the economic model, for which hazard estimates are similar.

Variation in ICERs across the three DRSMs (£122,328 to £142,825) was slightly greater than variation between the ERG approach (£124,807) and the company submission with a cap (£139,958). Advantages of using dynamic models in preference to the ERG approach have already been noted: dynamic models use all the data and avoid the arbitrary choice of which data to use for the extrapolating model. Two additional advantages of the DRSMs are first that model choice may be guided by clinical input into the likely behaviour of the long-term excess hazard, and secondly that external evidence is formally included as part of the model fitting procedure, instead of via a post-hoc adjustment. Collectively, this allows for a stronger emphasis on understanding the likely behaviour of the long-term excess hazard function and the plausibility of different assumptions about the long-term behaviour.

As a side-note, whilst estimates of effectiveness were similar in the replicated company approach (see Table 6.4), the replicated ERG approach provided markedly different estimates: QALYs gained were 0.66 and 0.34 for nivolumab and docetaxel (respectively) in this replication, whilst they were 0.89 and 0.44 (respectively) in the original ERG approach. A potential explanation is that the ERG's methodology may have been inappropriate; when critiquing it the NICE decision support unit noted that it was unclear if the ERG accounted for differing sample sizes when fitting the exponential model in Excel [223].

Table 6.5: Cost-effectiveness estimates from different extrapolation approaches.

	Absolute Value		**Incremental values**		**ICER**
	QALYs	**Cost**	**QALYs**	**Cost**	**(per QALY)**
Replicated submission (no cap)					
Nivolumab	1.29	£85,882	0.74	£65,470	£87,926
Docetaxel	0.55	£20,413			
Replicated submission (with cap)					
Nivolumab	0.95	£72,943	0.39	£54,412	£139,958
Docetaxel	0.56	£18,530			
Replicated ERG approach					
Nivolumab	0.66	£56,985	0.33	£40,799	£124,807
Docetaxel	0.34	£16,186			
Dynamic survival models					
Local trend					
Nivolumab	1.06	£75,060	0.50	£56,699	£113,170
Docetaxel	0.56	£18,361			
Damped trend					
Nivolumab	0.87	£67,328	0.35	£49,600	£141,236
Docetaxel	0.52	£17,728			
Dynamic relative survival models					
Local level					
Nivolumab	0.88	£67,880	0.36	£50,229	£139,657
Docetaxel	0.52	£17,651			
Local trend					
Nivolumab	0.99	£72,990	0.45	£54,847	£122,328
Docetaxel	0.54	£18,143			
Damped trend					
Nivolumab	0.86	£66,899	0.34	£49,196	£142,825
Docetaxel	0.52	£17,702			

DSM: dynamic survival model. ERG = Evidence review group. ICER = incremental cost-effectiveness ratio = incremental costs / incremental QALYs. QALY = quality-adjusted life-years.

6.4 Additional analyses

The analyses so far have demonstrated a number of potential benefits of dynamic models over current practice models. This includes an explicit modelling of the long-term trend, incorporating external data to inform extrapolations, and encoding clinical views on long-term survival via model specification. This section demonstrates further advantages of dynamic models: the ability to incorporate time-varying treatment effects and quantify extrapolation uncertainty. These two further topics are very important; a thorough treatment is beyond the scope of this thesis, but the fundamental points are introduced here with suggestions for further research.

6.4.1 Time-varying treatment effects

This case-study has focused on modelling the hazard function (or the excess hazard function) for the docetaxel group. In the original company submission, estimates for the nivolumab group were obtained by applying a single (time-invariant) hazard ratio to the docetaxel estimates. This hazard ratio represents the treatment effect, which was assumed to be constant over-time. A strength of DSMs is that it is straightforward to incorporate a time-varying treatment effect, relaxing the potentially restrictive assumption of a constant treatment effect. This is achieved in the same way that estimates of the level or trend of the hazard function are allowed to vary; by specifying a time series model for the temporal evolution. To illustrate this approach, the DRSM of Equation (3.3) in Chapter 3 is extended to include a time-varying excess hazard ratio, which in its general form is:

$$\begin{aligned}
Y_{t_i}^{(j)} &\sim \text{Poisson}\left([\exp(\beta_{1,t_i} + \delta_{1,t_i} \times j) + \lambda_{t_i}^{P}] \times \tau_{t_i}\right) \\
\beta_{1,t_i} &= \beta_{1,t_{i-1}} + \phi_1 \times \beta_{2,t_{i-1}} \times \omega_{t_i} \\
\beta_{2,t_i} &= \phi_1 \times \beta_{2,t_{i-1}} + \zeta_{t_i}^{\beta_2} \\
\delta_{1,t_i} &= \delta_{1,t_{i-1}} + \phi_2 \times \delta_{2,t_{i-1}} \times \omega_{t_i} \\
\delta_{2,t_i} &= \phi_2 \times \delta_{2,t_{i-1}} + \zeta_{t_i}^{\delta_2} \\
\zeta_{t_i}^{\beta_2} &\sim N(0, Z) \\
\zeta_{t_i}^{\delta_2} &\sim N(0, Z).
\end{aligned} \tag{6.1}$$

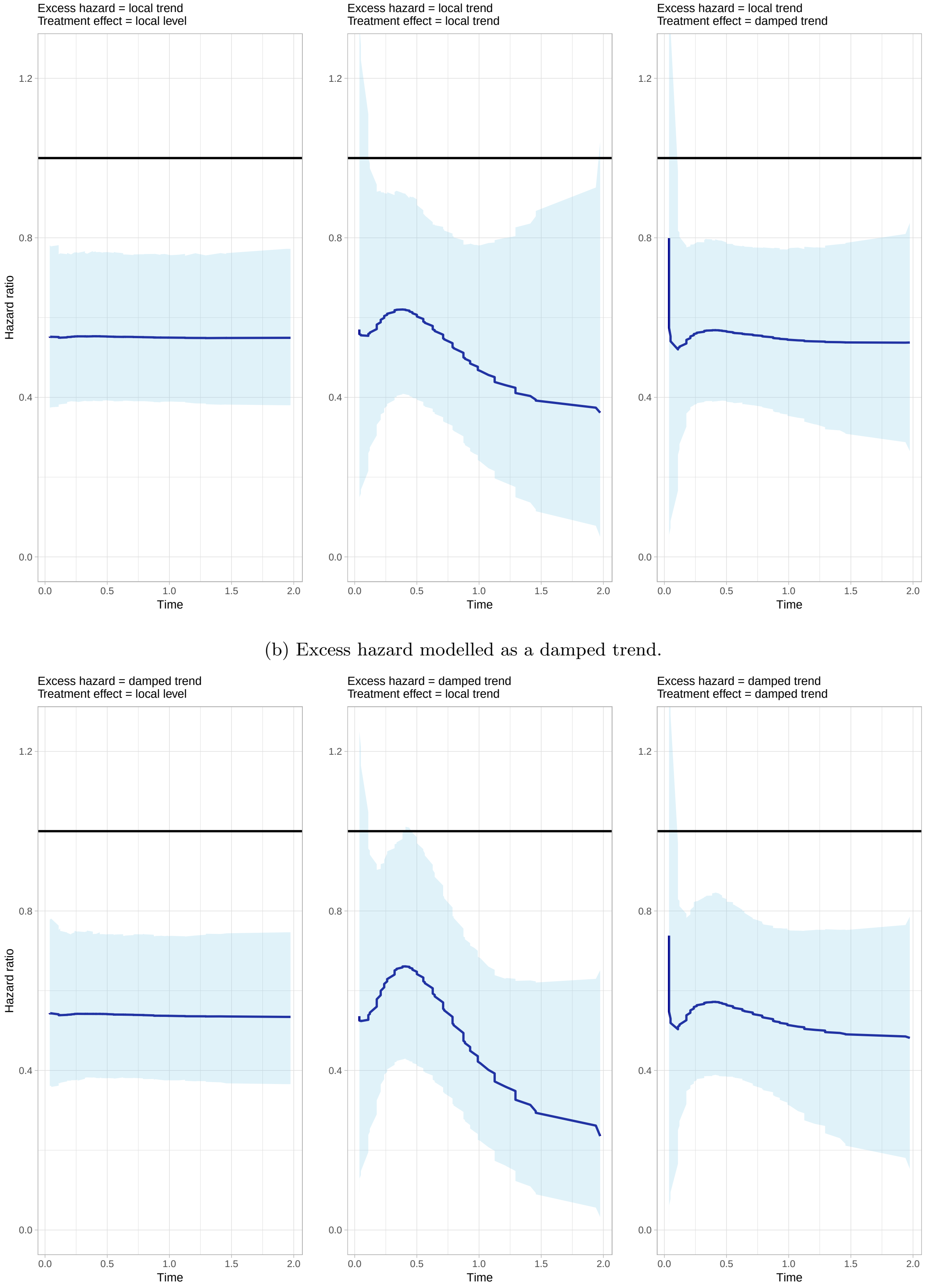
Where j is the group indicator (here $= 0$ for those receiving docetaxel and 1 for those receiving nivolumab), $Y^{(j)}$ are the observed deaths for group j, δ_{1,t_i} and δ_{2,t_i} are the level and trend for the treatment effect, and the trend has an innovation variances $\zeta_{t_i}^{\delta_2}$ (the innovation variance for β_{2,t_i} has similarly been amended). As with the excess hazard function, the treatment effect may be modelled as following either a local level, local trend, or damped trend time series model. Extensions to allow both the level and trend components of the treatment effect to vary over time are also possible, which would further increase the number of candidate models. As with standard DSMs, it is noted that the primary focus is on identifying and extrapolating any changes in the trend over time. Hence for this analysis all models with a trend have a global level.

This model requires specification of the time series models to use for both the excess hazard and the treatment effect. Six possible models are included here: three models for the treatment effect (local level, local trend, and damped trend) are displayed, whilst both a local trend and a damped trend are evaluated for the excess hazard in Figure 6.5 (a local level for the excess hazard was not considered as this visually gave a poor fit to the data in Figure 6.3). Both figures also include a black reference line for a hazard ratio of one (no treatment effect).

Visually, estimates of the nivolumab treatment effect were similar for the two approaches to modelling the excess hazard, but varied depending on the model used for the treatment effect. Use of a local level model for the treatment effect provided estimates that were virtually constant over time, whilst use of a local trend suggests a short-term decrease in the treatment effect, followed by a long-term increase (moving away from a hazard ratio of one). Estimates of uncertainty are also greater from the local trend model when compared with the local level model. As expected, use of a damped trend model for the treatment effect leads to estimates (both point estimates and estimates of uncertainty) that fall between the local level and local trend models. It is further noted that in Equation (6.1) the treatment effect is applied to the disease-specific (excess) hazard function. This is intuitively appealing, as it is likely to be more realistic than assuming that the treatment effect applies to the overall all-cause hazard. Modelling the treatment effect as applying to the overall hazard can lead to biased results as it includes the unrealistic assumption that treatment will reduce mortality that is unrelated to the disease [224]. Extending DRSMs to incorporate time-varying treatment effects allows for the analysis of both treatment groups simultaneously whilst also incorporating external data and employing very weak structural assumptions about both the baseline hazard and any treatment effect. As such, these models are very appealing. A potential drawback is the large number of possible model specifications. This could lead to cherry-picking of the specific models based on the extrapolations that they provide. Future work could identify how robust extrapolations are to model specification and how accurately these models can uncover the true treatment effect. This would require datasets with long-term observed treatment effects, which may be used as case-studies or used to inform the design of simulation studies. As no such datasets were identified during this fellowship, this was not pursued here.

Figure 6.5: Within-sample estimates of the treatment effect, with 95% confidence intervals.

(a) Excess hazard modelled as a local trend.



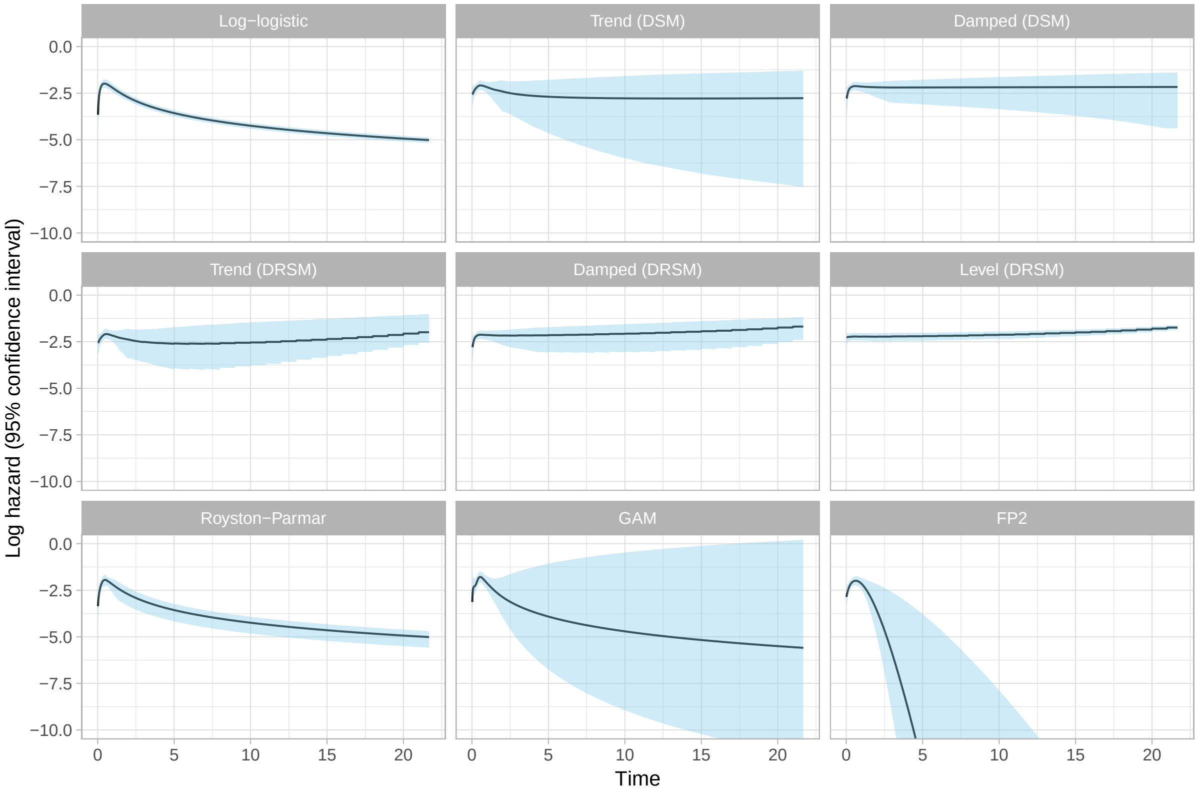
(b) Excess hazard modelled as a damped trend.

6.4.2 Extrapolation uncertainty

Estimates of the extrapolated hazard function will be associated with uncertainty. This uncertainty is incorporated within probabilistic sensitivity analyses and will impact on both decision uncertainty and value of information analyses (such as the expected value of collecting further information) [45]. A full treatment of these important topics is beyond the scope of this thesis, but would be worthy of future research. To illustrate this issue, the uncertainty in extrapolations arising from the models of this chapter is presented in Figure 6.6. This expands on work performed for this fellowship into extrapolation uncertainty [2]. This work considered the use of current practice models for extrapolation, using simulated representative hazard patterns and a hypothetical cost-effectiveness analysis. It found that estimates of uncertainty differed with model choice, with confidence interval widths varying three-fold. Further, uncertainty estimates were too narrow: the impact of model choice on decision uncertainty was marked, but extrapolation uncertainty had a near-negligible contribution to value of information estimates. In addition, the intuitive understanding that uncertainty should increase as the time-horizon of extrapolations increases was not reflected in estimates of uncertainty from many current practice models.

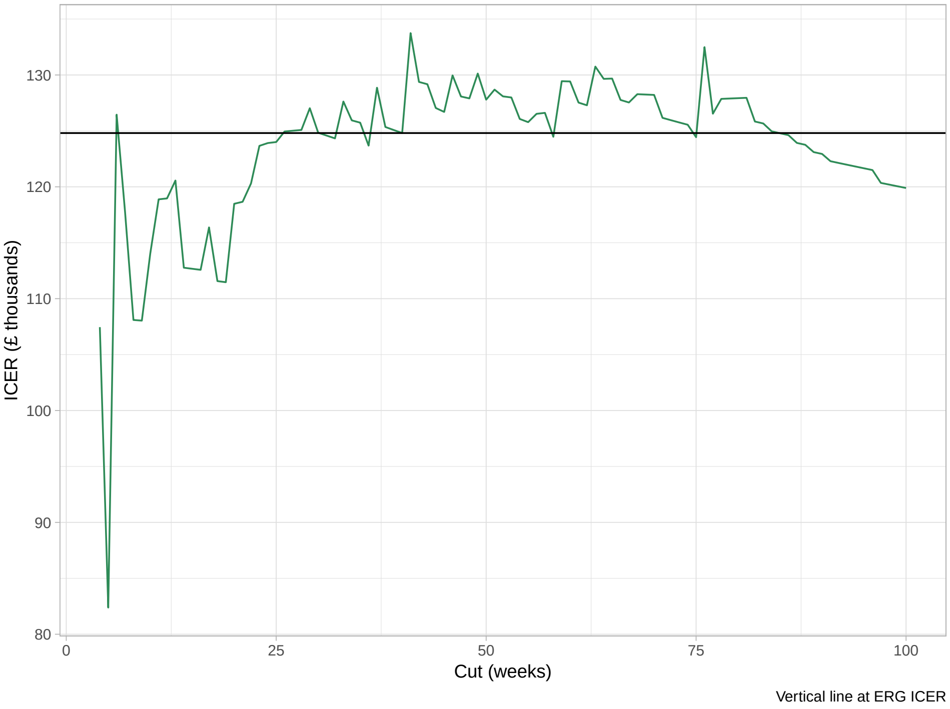
For Figure 6.6 uncertainty is quantified via 95% confidence intervals (or 95% credible intervals for the dynamic models, due to their Bayesian implementation) and includes GAMs and FPs for completeness. The current practice model (log-logistic), RPM and GAM all provide very similar point-estimates, but markedly different uncertainty estimates. Estimates of uncertainty do not fan-out over time for either the log-logistic or the RPM, whilst those from the GAM do. Uncertainty estimates arising from the DSMs also fan-out over time, with larger uncertainty arising from the local trend model than for the damped trend model. In general estimates of uncertainty from a damped trend model will be less than those from a corresponding non-damped DSM, due to the process of dampening. Both DSMs provide estimates of extrapolation uncertainty that are less wide than corresponding estimates from the GAM. For example, at 15 years the width of the 95% confidence interval from the GAM is approximately 10 units, whilst the width from the damped trend DSM is approximately 2.5 units (four times smaller). The two DRSMs demonstrate that by incorporating external evidence they can substantially reduce extrapolation uncertainty (as there is no modelled uncertainty for the general population lifetables). Extrapolations (point-estimates) from the best-fitting FP were implausible as they very quickly went to zero (so tended to $-\infty$ on the log-scale), although uncertainty estimates did fan-out over time. As demonstrated by Figure 6.6, estimates of uncertainty vary markedly by survival model. Further research could explore which model(s) provide realistic estimates of uncertainty. For example, whilst those from the log-logistic model appear to be too narrow in this example, it is unclear if the true magnitude of uncertainty is best reflected by the GAM, the damped trend DSM, or a different model. Ideally this future research would use real datasets, as simulation studies would be strongly affected by the choice of data-generating mechanism.

Figure 6.6: Model estimates of the hazard, with 95% confidence intervals.



Note that the ERG approach is not included in Figure 6.6, as there is no established method to model uncertainty in the choice of cut-point. The sensitivity of cost-effectiveness estimates to the choice of cut-point is illustrated in Figure 6.7; ICER values range from £82,000 to £134,000 depending on how much evidence is used in the extrapolation model. It is further noted that the use of non-parametric hazard estimates (as occurs for the ERG approach) can lead to a substantial increase in uncertainty when compared with a parametric model [27].

Figure 6.7: Sensitivity of the ICER to different cut values (ERG extrapolation approach).



6.5 Conclusions

This case-study has demonstrated that dynamic models can provide flexible fits to the observed data. A specific advantage of these models in contrast to other survival models is that they explicitly model a trend in the hazard, allowing for insight into the evolution of the hazard function over time. At the first NICE appraisal committee, there was disagreement between the company and ERG about the properties of the extrapolated hazard. The company believed that it was decreasing, whilst the ERG favoured a constant hazard (no trend). The results shown in Figure 1 demonstrate that the best estimate is a decreasing trend at the end of follow-up, but also that this estimate is associated with a lot of uncertainty and a constant trend cannot be ruled-out.

A further advantage of the DSMs is that, as local models, they can incorporate all the data without letting early outcomes unduly influence estimates towards the end of follow-up, as these will be more relevant when forming extrapolations. This is demonstrated in the trend analysis of Figure 1 where it is shown that at early times the trend is positive, but over time this becomes negative. This approach is more efficient than the approach of the ERG which was to discard all data collected before nine months when generating extrapolations.

The clinical plausibility of extrapolations is very important. Implausible extrapolations resulted from the use of both standard and spline-based survival models, as over time the hazard function fell below that of the age-matched general population. Similar extrapolations were seen for the local trend DSM. The extension to incorporate this external general population evidence and instead extrapolate the excess hazard (or relative survival) leads to DRSMs, which ensure that the extrapolated hazard function does not fall below that of the general population. Assessing if extrapolated hazards fall below corresponding general population hazards is not the only measure of extrapolation plausibility, but it is an important one that should be considered. Different model specifications are possible for DRSMs, reflecting different assumptions about the long-term behaviour of the excess hazard. The options included here were that the excess hazard was constant, the observed trend continued until the excess hazard became zero, or the observed trend continued in the short-term, with subsequent values of the excess hazard being constant. This flexibility in model specification and the direct interpretation of the extrapolations is a significant advantage of DRSMs when compared with other survival models and allows for the natural inclusion of clinical opinion about both the natural history of the disease and the likely mechanism of action of treatments. This case-study also demonstrated two additional important advantages of dynamic models. The first is the ease with which time-varying treatment effects may be incorporated, the second is that explicit modelling of innovation variances (the variation of model parameters over time) provides plausible estimates of extrapolation uncertainty.

Recreated patient-level data was used in this case-study. The recreated company submission demonstrated close agreement with the original submission, demonstrating the usefulness of using recreated data. One limitation with not having patient-level data is that it was not possible to explore the effects of covariates on survival. In particular, when estimating relative survival it has been demonstrated that including age can lead to increased accuracy [225]. It is straightforward to extend DRSMs to include age; this was implemented for the case-study of Section 4.1.

The results from the DRSMs suggest that the ICER arising from the company's original approach (£88,000) is likely to be too low; depending on the long-term prognosis of patients the ICER is likely to be between £122,000 and £143,000. This range of ICERs is above the acceptable threshold for end-of life-treatment, which is typically assumed to be £50,000. Following their original submission, the company offered a discount to the cost of their treatment to lower the ICER (and so improve the possibility of a positive recommendation). The magnitude of discount required to make the treatment cost-effective will be strongly affected by the extrapolation approach used. Of the approaches evaluated here, it is not possible to definitively state which extrapolation approach would be the preferred base-case analysis, but the use of a dynamic model which incorporates external evidence appears to be the most useful. Future research could identify the situations when the different DRSM specifications (including the modelling of the treatment effect) are appropriate.

Chapter 7

Discussion

Extrapolation of survival data plays a key role in HTA, influencing decisions about which technologies are funded and hence are made available to patients. The primary aim of this thesis is to provide answers to the question “What is the value and role of time series methods for generating predictions of future survival in HTAs?”. This chapter begins with an overview and discussion of the main findings from this thesis, discussing their implications along with how they add value to the current evidence on extrapolation in HTA. These results are consolidated within a suggested good practice framework for extrapolation in HTA, given the current evidence base. Limitations with the thesis are then discussed, which motivates areas for future research. A summary conclusion ends this chapter and the thesis.

7.1 Overview of key findings

Dynamic models were the main focus of this thesis, but a number of alternative survival models were also considered. As such, an evaluation of the role and value of dynamic models is contextualised by also considering the role and value of the alternative models.

In both simulation studies (Chapters 4 and 5) the dynamic models often provided some of the best overall goodness of fit. This was despite these models never being correctly specified. Similar results were observed for the GAM (Chapter 4) and RP cure model (Chapter 5). This suggests that sufficiently flexible models may be able to adequately describe and extrapolate complex hazard patterns. These more flexible models have less strict model assumptions than current practice models as they replace strong parametric assumptions with an assumption of smoothness. These weaker parametric assumptions may explain their robustness to misspecification. However, this did not always occur. For example, FPs and RPMs (non-cure) provided poor extrapolations, whilst all the models considered (flexible and current practice) provided poor extrapolations in scenarios with short follow-up where the turning point in the hazard function was not captured by any model. In addition, in the cure simulation, use of models without a cure fraction led to substantially worse extrapolations when compared to models with a cure fraction. In a similar vein, use of models with a cure fraction provided relatively poor extrapolations when the truth was a non-cure mixture model (and so not a special case of a cure model with a cure fraction of 0%).

These findings suggest that flexible models (such as DSMs, GAMs, DRSMs, and DCFMs) may be viable options for the analysis and extrapolation of survival data under certain circumstances. Namely, if there is adequate follow-up and if the presence of a cured fraction is appropriately identified and modelled. In practice it will be very difficult to determine if either of these circumstances are met. Having adequate follow-up will occur only if the observed data can be used to predict the unobserved future. This is by definition an assumption that cannot be verified. Similarly, for some disease areas, the presence of a cured fraction may only be known when patients have been followed-up to death. Results from the abiraterone case-study (Section 4.1) highlight the poor extrapolations that can occur in practice when there is insufficient follow-up. These considerations have two implications. The first is the importance of clinical input into both the disease natural history and the impact of treatment on this. This may provide some insight into any expected changes in the hazard function in the future and into the plausibility of assuming the presence of a cured fraction. The second implication is the need for additional (external) evidence or longer follow-up, where possible, to help inform extrapolations. For both of these implications there are advantages to using a dynamic model. First, clinical input into future outcomes can be reflected by the choice of model as demonstrated in the HTA case-study of Chapter 6, where the extrapolated excess (disease-specific) mortality could be constant or decrease to either zero or a non-zero number. Secondly, the two novel extensions of DSMs in Chapter 3 demonstrate that it is possible to incorporate external evidence within dynamic models. Increasing the length of follow-up can also produce improved extrapolations for dynamic models. For NICE appraisals of cancer drugs there is the possibility of interim funding via the cancer drugs fund. This is considered for appraisals in which the cost-effectiveness model is considered to be robust for decision making and the drug has the potential to be cost-effective but further data collection is required to reduce the uncertainty in the funding decision. This further data collection would result in longer follow-up times, hence increasing the relevance of dynamic models.

By incorporating time series methods, DSM can use of all the data when generating extrapolations whilst simultaneously giving more weight to more recent observations. The amount of additional weight to give is estimated from the data. This avoids the subjective choice of how much data to use in the extrapolating model, as estimates of cost-effectiveness can be sensitive to this choice. This is illustrated in Figure 6.7 of Chapter 6. Giving more weight to more recent observations increases the potential for DSMs to identify turning points in the hazard function near the end of follow-up. In the simulation study of Chapter 4 only DSMs and GAMs were able to correctly identify the long-term increases in the hazard (albeit, only in scenarios with long follow-up). A further advantage of incorporating time series methods is that extrapolation uncertainty from these models reflects the intuitive notion that the more distant the prediction the more uncertain it is. In addition, the use of Bayesian methods for model estimation provides a natural representation of parameter uncertainty (via posterior distributions) for use in cost-effectiveness analyses [226].

Both dynamic models and emerging practice models provide flexible within-sample estimates. Dynamic models also provide parsimonious extrapolating models which have direct interpretations. For example, the local trend model used in this thesis assumes that the extrapolated log-hazard will follow a linear model. As such, it allows for an explicit quantification of the distribution of the extrapolated trend. This allows for questions such as if the extrapolated trend is positive or negative, and if the uncertainty in this trend includes zero. These are important questions for an appraisal committee, which alternative models cannot easily answer, as discussed in Chapter 6. The damped trend DSM allows for an extrapolated trend that decreases as extrapolations move further into the future so that eventually extrapolations become constant.

This is a particularly attractive assumption as it combines the desire to incorporate observed evidence about the trend at the end of follow-up with the desire for conservative extrapolations (to avoid implausible estimates). In the mixture-Weibull simulation study of Chapter 4 damped trend models generally out-performed local trend models, whilst giving near-identical performance (of the cure fraction extensions) in the cure fraction simulation study of Chapter 5. For both simulation studies all three versions of the damped trend model (DSM, DCFM, DRSM) provided some of the best overall goodness of fit values. Dampening of the extrapolated trend also helps to alleviate the impact of an unexpected turning-point in the hazard function during the extrapolated period. This occurred in the abiraterone case-study in Section 4.1, where damped trend models provided the best extrapolations. The use of damped trend models to extrapolate survival data is a novel development for this thesis. This includes creating an algorithm to accommodate unevenly spaced observations.

Dynamic models can also be extended to incorporate time-varying treatment effects. This relaxes the restrictive assumption that treatment effects are constant over time. Again, the use of explicit models for the extrapolated treatment effect enables both inferences (such as if the uncertainty in the extrapolations include no treatment effect) and scenario analyses around different assumptions. This includes a local level model for no trend in treatment effects and a damped trend model for changes in treatment effects that persist in the short-term but not the long-term. This extension can be combined with a DRSM to provide a single analytic framework for combining short-term trial data with longer-term external data, whilst explicitly modelling the evolution of the disease-specific hazard function and any time-varying effects of treatment on this. It is noted that the 2013 NICE ‘Guide to the methods of technology appraisal’ recommends assessing three possibilities for extrapolations: a continued effect, a diminishing effect, and no further effect beyond the end of follow-up [10]. These assumptions relate to local trend, damped trend, and local level dynamic models for the treatment effect, respectively. Finally, the use of dynamic models requires the very weak structural assumption that outcomes which are close together in time will exhibit some degree of correlation, with the magnitude of correlation estimated from the data. This structural assumption is much less restrictive than the assumptions employed by current practice models (including their extension to include a cure fraction). The results of Chapter 5 demonstrate that dynamic models are more robust to model misspecification than current practice models.

7.2 Contributions to the literature

The analyses of this thesis represent the first applications of DSMs in HTA. The extrapolation of survival outcomes is an important and common task in HTA, but less frequent in other disciplines. As such, there was little existing literature on the use of DSMs for extrapolation, with studies typically focusing on obtaining flexible within-sample estimates. The results of this thesis provide useful evidence on model specification when using DSMs for extrapolation, including the choice of prior distribution for innovation variances and how to generate continuous extrapolations for a damped trend model. Two novel extensions to the DSMs were also developed to incorporate external evidence in the form of life tables. Both relative survival and cure fraction methods were included. The resulting models are of particular importance for extrapolations as the inclusion of long-term mortality data helps to simultaneously reduce the uncertainty in extrapolations and improve their plausibility. These extensions to dynamic models also benefit from the direct interpretation of the extrapolating model. The ability to base model choice on the likely long-term hazard (or excess hazard) function is of particular benefit since the results of this thesis demonstrate that measures of within-sample goodness of fit have very limited use when choosing an extrapolating model. The analysis of current practice models in Section 4.3.1 provides a dramatic example of this as the Gompertz nearly always had the best within-sample fit but the worst extrapolations.

Any assessment of the value of dynamic models will depend on the comparator models that are included. The field of survival analysis is extensive, so comprehensive literature reviews were performed in Chapter 2 to ensure that appropriate comparators were identified. These reviews covered both the academic literature (peer-reviewed and grey literature) as well as NICE appraisals. The multitude of evidence sources allowed for the categorisation of methods into ‘current practice’ which served as a benchmark when evaluating model goodness of fit, and ‘emerging practice’ which included other models that may be used for extrapolation in HTA and so represent viable alternatives.

Simulation studies, covering a wide variety of scenarios and models, were conducted in Chapters 4 and 5 to evaluate the within-sample fit and extrapolation performance of DSMs and comparator models. This included various model specifications, differing lengths of follow-up and sample size, and two data-generating mechanisms. These mechanisms were a mixture-Weibull and a Weibull cure model. Both represent clinically plausible scenarios of heterogeneous outcomes, due to either patient frailty or a cured fraction. These simulation studies were complemented by a case-study using patient-level data and a re-analysis of an existing NICE technology appraisal using recreated (via digitisation) patient-level data. The appraisal re-analysis of Chapter 6 represents the first use of DSMs and DRSMs in HTA and demonstrated their use in this setting. Together, these analyses illustrate the feasibility of estimating dynamic models using state of the art (HMC) Bayesian methods.

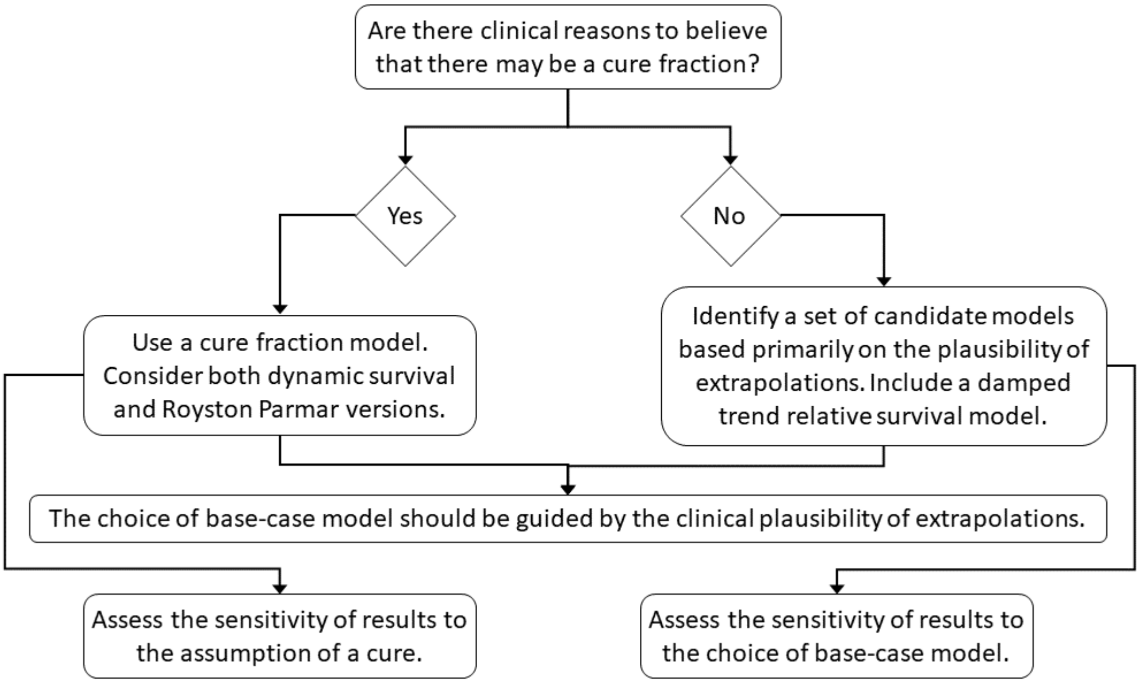
The mixture (non-cure) simulation study of Chapter 4 was relatively simple, comprising two monotonic (Weibull) hazard functions. Yet producing accurate extrapolations, with low bias and low variance, remained challenging for the majority of methods, even with a follow-up of three years. Across the nine scenarios considered the current practice models had some of the best overall goodness of fit values, despite entirely failing to capture the long-term shape of the hazard function. This dataset and the results of this thesis will provide useful test-cases and benchmarks for future research to see if it is possible to improve on the current extrapolation performance of both current practice models and the dynamic models. A further contribution to the literature is the finding that when multiple model specifications are available, models that include the logarithm of time as a covariate provide superior extrapolations to models that include time without a transformation. This issue has not been discussed previously in the literature to my knowledge. Both approaches provide similar within-sample estimates but use of the logarithm of time contracts the time variable and so reduces the impact of extrapolating an incorrect trend.

Collectively, the analyses of this thesis provide valuable insight into the performance of dynamic models along with current and emerging practice in different situations. These findings are expanded on and encapsulated within the good practice guidance contained in the next section.

7.3 Good practice guidance on the role of dynamic models

A flowchart to assist in model choice is provided in Figure 7.1. Additional considerations are also listed to support this flowchart.

Figure 7.1: Flowchart for model selection.


Are there clinical reasons to believe that there may be a cure fraction?
Yes
No
Use a cure fraction model. Consider both dynamic survival and Royston Parmar versions.
Identify a set of candidate models based primarily on the plausibility of extrapolations. Include a damped trend relative survival model.
The choice of base-case model should be guided by the clinical plausibility of extrapolations.
Assess the sensitivity of results to the assumption of a cure.
Assess the sensitivity of results to the choice of base-case model.

The analyses of this thesis suggest that when the truth includes a cure fraction then cure models will provide good estimates for both the within sample and extrapolated period. As such, this should be a key consideration in model choice. Of the cure models considered, the dynamic and RPMs appear to be the most robust to misspecification of the hazard function for the uncured fraction and so both should be considered. When there is uncertainty about the presence of a cure fraction the sensitivity of results to the use of non-cure models should be considered. When there was no cure fraction none of the models considered in this thesis (current and emerging practice, including dynamic models) performed consistently well, highlighting the importance of considering multiple candidate models. Flexible models (DSMs and emerging practice) can fit better the observed data and so extrapolate trends observed towards the end of follow-up. However, they are also at danger of overfitting and so extrapolating non-existent trends. This danger increased with shorter follow up and/or smaller sample size. Hence current practice models, which were less flexible and so in general less likely to generate implausible extrapolations, should also be considered as candidate extrapolation models. FPs were particularly vulnerable to overfitting as were models that didn't use the logarithm of time (including the Gompertz), so neither should be considered. A damped trend model partly mitigates the danger of extrapolating an incorrect trend, so should be considered as a candidate model. Incorporating external data helps to improve the plausibility of extrapolations whilst reducing their variability. In the absence of a cured fraction a damped trend DRSM appears to outperform the other methods for including external evidence considered here. As such, it is recommended that a damped trend DRSM be included in the set of candidate extrapolation methods. The choice of base-case model should be driven by considerations of clinical plausibility, noting also that within sample fit can be a poor predictor of extrapolation performance. To aid this, it is very useful to plot the empirical hazard function. Benefits of this include the direct clinical interpretation (as it is non-cumulative, it is more informative than the cumulative hazard function), and hazard estimates are very similar to the transition probabilities that are used in health economic models. As part of this fellowship, code has been made available to generate hazard plots [2]. Sensitivity to model choice should always be assessed.

7.4 Limitations and future research

The simulation studies of Chapters 4 and 5 considered a wide variety of models for data with and without a cure fraction. Whilst the results for dynamic models were promising, no single model was identified as providing the best estimates across the range of scenarios considered. To compound this, the use of a model without (with) a cure fraction led to poor extrapolations when the true data-generating mechanism included (did not include) a cure fraction, for the data-generating mechanisms considered. In addition, the simulation studies only looked at two data-generating mechanisms and only three case-studies were included. Hence more experience is required of the performance of DSMs, their extensions, and comparator models. This would ideally include simulation studies with different data-generating mechanisms and further case-studies, including in non-cancer disease areas. The three case-studies were for different diseases (prostate cancer, non-small-cell lung cancer, and breast cancer) and all demonstrated an initial increase from a low value, followed by a turning point and a long-term decrease back to a low value. Many studies do not plot the hazard function, but similar hazard patterns have also been observed for Merkel cell carcinoma and melanoma [209, 227]. This raises the interesting potential that the shape of hazard function observed in clinical oncology trials follows a common pattern. Further quantification of hazard patterns, including for non-cancer trials, and an assessment of their impact on extrapolations when included as external evidence, is likely to be a key future research area. I am currently supervising a PhD short-project into using external trial evidence to inform extrapolations (using the case-study of Section 4.1).

The specification of dynamic models followed that used in the literature ([128, 143]) which was Bayesian, whilst comparator models used available software ([100, 140]) and were frequentist in implementation. Ideally these two approaches would not be compared due to their different assumptions. The Bayesian approach assumes that parameters are random and the sample fixed. For the frequentist approach this assumption is the other way around. Some of the observed findings may be due to these differences in approach. For example, the performance of GAMs was often similar to DSMs; of the frequentist models the GAMs are the only ones to penalise for model complexity during model fitting, which makes them the most similar to a Bayesian approach (as discussed in Chapter 4 of Wood [135]). The current comparisons are motivated by noting that Bayesian versions of current practice are not routinely used (although implementations are available [220]), whilst Bayesian implementations of emerging practice models were not available at the time of this thesis.

A comparison of the extrapolation performance of DSMs and GAMs was conducted for both the Weibull-mixture simulation study (Chapter 4) and the abiraterone case-study (Section 4.1). The results suggest that both models may be useful for the extrapolation of survival data. This finding for DSMs supports the hypothesis of this thesis that time series methods are useful for the extrapolation of survival data, as they take advantage of the temporal structure of the data. The finding for GAMs was not anticipated, further research into the reasons for this relatively good performance would be beneficial. Also, it would be useful to replicate the extensions to DSMs (such as to incorporate external evidence) for GAMs. The penalty term in GAMs ensures a certain degree of smoothness in estimates; this implicitly acknowledges the temporal structure of data, as smoothness requires that estimates which are close in time are similar.

For the DSM specifications considered in this thesis up to five model types are possible (local level, local trend with a global or local level, damped trend with a global or local level). In addition, novel extensions to incorporate external evidence resulted in DCFMs and DRSMs. A further extension to incorporate time-varying treatment effects was also demonstrated. It would be possible to combine all these extensions, providing a DCFM in which the hazard for the uncured fraction follows a DRSM with a treatment effect acting upon the excess hazard for the uncured population and/or the cure fraction. A (potentially time-varying) standardised mortality ratio could also be applied to the general population mortality data for the cured fraction if it is believed that the 'cured' people still have an elevated risk compared to the general population. Such a model is likely to more closely reflect reality than any of the models currently considered (either in this thesis or in the literature), but it also presents an extensive range of potential model specifications. This raises the risk of potentially cherry-picking the model that provides the most convenient extrapolations for the user. Hence future research should explore the robustness of extrapolations to model specification along with identifying the situations when certain specifications will be more useful. An alternative approach would be to consider averaging results from different model specifications. I am supervising an MSc dissertation exploring this using the case-study of Chapter 6.

There are a variety of other options for expanding the research of this thesis. These include:

- Extrapolation uncertainty and time-varying effects were briefly considered in this thesis. A thorough exploration of these topics was beyond the scope of this thesis and should be covered in future research.

- DSMs have not previously been used in HTA. There is a learning-curve associated with the use of a new statistical method and this opportunity cost has not been considered. That is, do the incremental benefits of using DSMs outweigh the cost of learning to use the models? Ideally, future developments would include user-friendly interfaces to fit DSMs and their extensions.

- The innovation variance controls the smoothness (or wiggliness) of the model-based hazard estimate over time. For the applications considered here a constant variance was assumed in order to decrease the potential for over-fitting that could occur if the model were too flexible. Other applications of DSMs have used time-varying innovation variances [40, 128]. This could be explored in future research; for example the innovation variation could be correlated with sample size; allowing for larger variation at the start of follow-up (where there are more data) and less at the end. This could be achieved by linking the innovation variance to the at-risk sample size.

- Comparing estimates from models fitted with HMC to those fitted using filtering methods, which can be based on minimising one-step ahead predictions.

7.5 Conclusions

Survival data describe the occurrence of deaths over time and so form a natural time series. This motivates the use of dynamic models, which can exploit the temporal evolution of the hazard function when generating extrapolations. These models combine flexible within-sample estimates with simple models for extrapolations which have meaningful clinical interpretations. Compared with the parametric models representing current practice, dynamic models make minimal structural assumptions. This makes them less prone to model misspecification, which can impact on goodness of fit. A further strength of dynamic models is that it is relatively straightforward to extend them to incorporate external evidence and time-varying treatment effects which increases their applicability when used for extrapolation in HTA. Despite these strengths of dynamic models, it should be noted that they also have limitations. The simulation studies demonstrated that, as with all the flexible models, they can provide poor extrapolations when there is short follow-up data or small sample sizes. Extrapolation performance can be improved by incorporating external evidence, which is an important area for future research. In addition, the abiraterone case-study demonstrated that if the available evidence can not be used to predict the future, for example if there is a turning point in the hazard during the extrapolated period, then predictions of the future will be poor. This limitation is applicable to all models that are used for extrapolation, again future research into the use of external evidence to alleviate this problem would be fruitful.

To conclude, dynamic models for the analysis and extrapolation of survival data have value in HTA. Currently, their role may be best thought of as complementary to both the current practice and spline-based models identified in this thesis. More experience of these models and their extensions when used with different datasets is required to provide more specific guidance about their role in HTA, including when they should be used as the base-case model. Dynamic models are another option in the toolkit of methods for the analysis and extrapolation of survival data. They are theoretically attractive, and the initial results of this thesis are promising, with dynamic models often providing some of the best within-sample estimates and extrapolations.

Bibliography

[1] Kearns B, Stevenson M, Triantafyllopoulos K, Manca A. Generalized Linear Models for Flexible Parametric Modeling of the Hazard Function [Journal Article]. Medical Decision Making. 2019;39(7):12. Available from: https://doi.org/10.1177/0272989X19873661https://journals.sagepub.com/doi/pdf/10.1177/0272989X19873661.

[2] Kearns B, Stevens J, Ren S, Brennan A. How Uncertain is the Survival Extrapolation? A Study of the Impact of Different Parametric Survival Models on Extrapolated Uncertainty About Hazard Functions, Lifetime Mean Survival and Cost Effectiveness [Journal Article]. PharmacoEconomics. 2019;38(2):1–12.

[3] Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices [Journal Article]. New England journal of medicine. 1977;296(13):716–721. Available from: https://www.nejm.org/doi/full/10.1056/NEJM197703312961304?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[4] Masic I, Miokovic M, Muhamedagic B. Evidence based medicine–new approaches and challenges [Journal Article]. Acta Informatica Medica. 2008;16(4):219. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3789163/pdf/aim-2008-16-4-10.pdf.

[5] Campbell M. A statistician on a NICE committee [Journal Article]. Significance. 2010;7(2):81–84.

[6] Kearns B, Pandor A, Stevenson M, Hamilton J, Chambers D, Clowes M, et al. Cabazitaxel for Hormone-Relapsed Metastatic Prostate Cancer Previously Treated With a Docetaxel-Containing Regimen: An Evidence Review Group Perspective of a NICE Single Technology Appraisal [Journal Article]. PharmacoEconomics. 2017;35(4):415–424. Available from: https://doi.org/10.1007/s40273-016-0457-1https://link.springer.com/article/10.1007%2Fs40273-016-0457-1.

[7] Kearns B, Chilcott J, Whyte S, Preston L, Sadler S. Cost-effectiveness of screening for ovarian cancer amongst postmenopausal women: a model-based economic evaluation [Journal Article]. BMC medicine. 2016;14(1):200. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5139096/pdf/12916_2016_Article_743.pdf.

[8] Kearns B, Rafia R, Leaviss J, Preston L, Brazier J, Palmer S, et al. The cost-effectiveness of changes to the care pathway used to identify depression and provide treatment amongst people with diabetes in England: a model-based economic evaluation [Journal Article]. BMC health services research. 2017;17(1):78. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5259945/pdf/12913_2017_Article_2003.pdf.

[9] Kearns B, Michaels J, Stevenson M, Thomas S. Cost-effectiveness analysis of enhancements to angioplasty for infrainguinal arterial disease [Journal Article]. British Journal of Surgery. 2013;100(9):1180–1188. Available from: https://bjssjournals.onlinelibrary.wiley.com/doi/abs/10.1002/bjs.9195.

[10] [Web Page]; 2013. Available from: https://www.nice.org.uk/process/pmg9/chapter/foreword.

[11] [Web Page]; 2017. Available from: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-highly-specialised-technologies-guidance/HST-interim-methods-process-guide-may-17.pdf.

[12] [Web Page]; 2017. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance.

[13] Patel R, Sweeting MJ, Powell JT, Greenhalgh RM, Investigators ET. Endovascular versus open repair of abdominal aortic aneurysm in 15-years' follow-up of the UK endovascular

aneurysm repair trial 1 (EVAR trial 1): a randomised controlled trial [Journal Article]. The Lancet. 2016;388(10058):2366–2374. Available from: https://spiral.imperial.ac.uk:8443/bitstream/10044/1/43146/2/PIIS0140673616311357.pdf.

[14] Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial [Journal Article]. The Lancet. 2015;.

[15] Benedict A, Muszbek N, Perampaladas K. Survival modeling in UK oncology technology appraisals since the publication of good practice guidelines. In: European Journal of Cancer. vol. 49. Elsevier Sci LTD;. p. S328–S328.

[16] Gallacher D, Auguste P, Connock M. How Do Pharmaceutical Companies Model Survival of Cancer Patients? A Review of NICE Single Technology Appraisals in 2017 [Journal Article]. International journal of technology assessment in health care. 2019;35(2):160–167.

[17] Collett D. Modelling survival data in medical research (Third Edition). CRC press; 2015.

[18] Latimer NR. Survival analysis for economic evaluations alongside clinical trials–extrapolation with patient-level data: inconsistencies, limitations, and a practical guide [Journal Article]. Med Decis Making. 2013;33(6):743–54. Available from: https://journals.sagepub.com/doi/pdf/10.1177/0272989X12472398.

[19] Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach [Journal Article]. Med Decis Making. 2014;34(3):343–51. Available from: https://journals.sagepub.com/doi/full/10.1177/0272989X13497998.

[20] Latimer NR. In: NICE Decision Support Unit Technical Support Documents. London: National Institute for Health and Care Excellence (NICE) unless otherwise stated. All rights reserved.; 2013. .

[21] Bell Gorrod H, Kearns B, Thokala P, Labeit A, Stevens J, Latimer N, et al. Plausible and consistent tails: a review of survival extrapolation methods used in technology appraisals of cancer treatments [Journal Article]. Med Decis Making. 2019;.

[22] Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information [Journal Article]. Med Decis Making. 2017;37(4):353–366. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27681990https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6190619/pdf/10.1177_0272989X16670604.pdf.

[23] Jackson CH, Sharples LD, Thompson SG. Survival models in health economic evaluations: balancing fit and parsimony to improve prediction [Journal Article]. The international journal of biostatistics. 2010;6(1).

[24] Therapeutics C. Pixantrone monotherapy for treating multiply relapsed or refractory aggressive non-Hodgkin's B-cell lymphoma; 2012.

[25] Grieve R, Hawkins N, Pennington M. Extrapolation of survival data in cost-effectiveness analyses: improving the current state of play [Journal Article]. Med Decis Making. 2013;33(6):740–2. Available from: https://journals.sagepub.com/doi/pdf/10.1177/0272989X13492018.

[26] Gelber RD, Goldhirsch A, Cole BF. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. International Breast Cancer Study Group [Journal Article]. Control Clin Trials. 1993;14(6):485–99. Available from: https://www.sciencedirect.com/science/article/abs/pii/019724569390029D?via%3Dihub.

[27] Kearns B, Jones ML, Stevenson M, Littlewood C. Cabazitaxel for the second-line treatment of metastatic hormone-refractory prostate cancer: a NICE single technology appraisal [Journal Article]. Pharmacoeconomics. 2013;31(6):479–488. Available from: https://link.springer.com/article/10.1007%2Fs40273-013-0050-9.

[28] Latimer NR. Response to "survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach" by

Bagust and Beale [Journal Article]. Med Decis Making. 2014;34(3):279–82. Available from: https://journals.sagepub.com/doi/full/10.1177/0272989X13511302?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[29] Booth H, Tickle L. Mortality modelling and forecasting: A review of methods [Journal Article]. Annals of actuarial science. 2008;3(1-2):3–43.

[30] Kontis V, Bennett JE, Mathers CD, Li G, Foreman K, Ezzati M. Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble [Journal Article]. The Lancet. 2017;389(10076):1323–1335. Available from: https://spiral.imperial.ac.uk:8443/bitstream/10044/1/41813/7/PIIS0140673616323819.pdf.

[31] Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture [Journal Article]. Statistics in medicine. 2017;36(22):3443–3460.

[32] Ramnath S, Rock S, Shane P. The financial analyst forecasting literature: A taxonomy with suggestions for further research [Journal Article]. International Journal of Forecasting. 2008;24(1):34–75.

[33] Weron R. Electricity price forecasting: A review of the state-of-the-art with a look into the future [Journal Article]. International journal of forecasting. 2014;30(4):1030–1081.

[34] Lin WT. Modeling and forecasting hospital patient movements: Univariate and multiple time series approaches [Journal Article]. International Journal of Forecasting. 1989;5(2):195–208.

[35] Meade N, Islam T. Forecasting in telecommunications and ICT—A review [Journal Article]. International Journal of Forecasting. 2015;31(4):1105–1126.

[36] Clark JS, Bjørnstad ON. Population time series: process variability, observation errors, missing values, lags, and hidden states [Journal Article]. Ecology. 2004;85(11):3140–3150.

[37] Chatfield C. The analysis of time series: an introduction. CRC press; 2016.

[38] Breslow N. Discussion of ”Regression models and life-tables” by DR Cox [Journal Article]. Journal of the Royal Statistical Society Series B-Methodological. 1972;34(2):216–217.

[39] Demarqui FN, Loschi RH, Dey DK, Colosimo EA. A class of dynamic piecewise exponential models with random time grid [Journal Article]. Journal of Statistical Planning and Inference. 2012;142(3):728–742. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0378375811003521?via%3Dihub.

[40] Hemming K, Shaw J. A class of parametric dynamic survival models [Journal Article]. Lifetime Data Analysis. 2005;11(1):81–98.

[41] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations [Journal Article]. European Journal of Epidemiology. 2016;31(4):337–350. Available from: https://doi.org/10.1007/s10654-016-0149-3https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/pdf/10654_2016_Article_149.pdf.

[42] Hyndman R, Koehler AB, Ord JK, Snyder RD. Forecasting with exponential smoothing: the state space approach. Springer Science and Business Media; 2008.

[43] Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra [Journal Article]. Health economics. 2005;14(4):339–347. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.985.

[44] Heath A, Manolopoulou I, Baio G. A review of methods for analysis of the expected value of information [Journal Article]. Medical decision making. 2017;37(7):747–758. Available from: https://journals.sagepub.com/doi/full/10.1177/0272989X17697692?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[45] Rothery C, Strong M, Koffijberg HE, Basu A, Ghabri S, Knies S, et al. Value of information analytical methods: report 2 of the ISPOR value of information analysis emerging good practices task force [Journal Article]. Value in Health. 2020;23(3):277–286.

[46] Sculpher M, Palmer S. After 20 Years of Using Economic Evaluation, Should NICE be Considered a Methods Innovator? [Journal Article]. PharmacoEconomics. 2020;p. 1–11.

[47] Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies [Journal Article]. Health Information and Libraries Journal. 2009;26(2):91–108. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1471-1842.2009.00848.x.

[48] James KL, Randall NP, Haddaway NR. A methodology for systematic mapping in environmental sciences [Journal Article]. Environmental evidence. 2016;5(1):7.

[49] Cooper ID. What is a "mapping study?" [Journal Article]. Journal of the Medical Library Association: JMLA. 2016;104(1):76. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722648/pdf/mlab-104-01-76.pdf.

[50] O'Cathain A, Thomas KJ, Drabble SJ, Rudolph A, Goode J, Hewison J. In: A systematic mapping review of published qualitative research undertaken with specific randomised controlled trials. NIHR Journals Library; 2014. .

[51] Ramer SL. Site-ation pearl growing: methods and librarianship history and theory [Journal Article]. Journal of the Medical Library Association. 2005;93(3):397. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1175807/pdf/i0025-7338-093-03-0397.pdf.

[52] Gentles SJ, Charles C, Nicholas DB, Ploeg J, McKibbon KA. Reviewing the research methods literature: principles and strategies illustrated by a systematic overview of sampling in qualitative research [Journal Article]. Systematic reviews. 2016;5(1):172. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059917/pdf/13643_2016_Article_343.pdf.

[53] Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P. The role of modelling in prioritising and planning clinical trials [Journal Article]. Health Technol Assess. 2003;7(23):1–125.

[54] [Web Page]; 2017. Available from: http://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=1&SID=W1fpiLluaL58YjFV9OF&page=1&doc=1.

[55] Billingham L, Abrams K. Simultaneous analysis of quality of life and survival data [Journal Article]. Statistical methods in medical research. 2002;11(1):25–48. Available from: `https://journals.sagepub.com/doi/abs/10.1191/0962280202sm269ra`.

[56] Latimer NR. The role of treatment crossover adjustment methods in the context of economic evaluation [PhD]; 2012. Available from: `http://etheses.whiterose.ac.uk/3720/`.

[57] Squires H. A methodological framework for developing the structure of Public Health economic models [PhD]; 2014. Available from: `http://etheses.whiterose.ac.uk/5316/`.

[58] Youn JH. Modelling Health and Healthcare for an Ageing Population [PhD]; 2016. Available from: `http://etheses.whiterose.ac.uk/13982/`.

[59] Allignol A Arthur; Latouche. CRAN Task View: Survival Analysis; 2016. Available from: `https://CRAN.R-project.org/view=Survival`.

[60] Booth A, Sutton A, Papaioannou D. Systematic approaches to a successful literature review. Sage; 2016.

[61] Hutton J, Ashcroft R. What does "systematic" mean for reviews of methods [Journal Article]. Health services research methods: a guide to best practice London: BMJ Publishing Group. 1998;p. 249–54.

[62] Edwards S, Lilford R, Kiauka S. Different types of systematic review in health services research [Journal Article]. Health services research methods: a guide to best practice. 1998;p. 255–259.

[63] Gallacher D, Auguste P, Connock M. How do pharmaceutical companies model survival of cancer patients? A review of NICE single technology appraisals in 2017 [Journal Article]. International journal of technology assessment in health care. 2019;35(2):160–167.

[64] Benaglia T, Jackson CH, Sharples LD. Survival extrapolation in the presence of cause specific hazards [Journal Article]. Stat Med. 2015;34(5):796–811. Available from: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4847642/pdf/SIM-34-796.pdf`.

[65] Chu PC, Wang JD, Hwang JS, Chang YY. Estimation of life expectancy and the expected years of life lost in patients with major cancers: extrapolation of survival curves under high-censored rates [Journal Article]. Value Health. 2008;11(7):1102–9.

[66] Day SM, Reynolds RJ, Kush SJ. Extrapolating published survival curves to obtain evidence-based estimates of life expectancy in cerebral palsy [Journal Article]. Dev Med Child Neurol. 2015;57(12):1105–18. Available from: `https://onlinelibrary.wiley.com/doi/full/10.1111/dmcn.12849`.

[67] Demiris N, Lunn D, Sharples LD. Survival extrapolation using the poly-Weibull model [Journal Article]. Stat Methods Med Res. 2015;24(2):287–301. Available from: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4456429/pdf/10.1177_0962280211419645.pdf`.

[68] Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies [Journal Article]. Stat Med. 2006;25(11):1960–75. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2366`.

[69] Gong Q, Fang L. Asymptotic properties of mean survival estimate based on the Kaplan-Meier curve with an extrapolated tail [Journal Article]. Pharm Stat. 2012;11(2):135–40. Available from: `https://deepblue.lib.umich.edu/bitstream/handle/2027.42/90597/pst514.pdf?sequence=1`.

[70] Hwang JS, Wang JD. Monte Carlo estimation of extrapolation of quality-adjusted survival for follow-up studies [Journal Article]. Stat Med. 1999;18(13):1627–40.

[71] Andersson TM, Dickman PW, Eloranta S, Lambe M, Lambert PC. Estimating the loss in expectation of life due to cancer using flexible parametric survival models [Journal Article]. Statistics in medicine. 2013;32(30):5286–5300.

[72] Haberman S, Renshaw A. Mortality, longevity and experiments with the Lee-Carter model [Journal Article]. Lifetime Data Anal. 2008;14(3):286–315. Available from: `https://link.springer.com/article/10.1007%2Fs10985-008-9084-2`.

[73] Majer IM, Stevens R, Nusselder WJ, Mackenbach JP, van Baal PH. Modeling and forecasting health expectancy: theoretical framework and application [Journal Article]. Demography. 2013;50(2):673–97. Available from: https://link.springer.com/article/10.1007%2Fs13524-012-0156-2.

[74] McNown R, Rogers A. Forecasting mortality: a parameterized time series approach [Journal Article]. Demography. 1989;26(4):645–60.

[75] Messori A, Trippoli S. A new method for expressing survival and life expectancy in lifetime cost-effectiveness studies that evaluate cancer patients (review) [Journal Article]. Oncol Rep. 1999;6(5):1135–1141. Available from: https://www.spandidos-publications.com/or/6/5/1135.

[76] Nelson CL, Sun JL, Tsiatis AA, Mark DB. Empirical estimation of life expectancy from large clinical trials: Use of left-truncated, right-censored survival analysis methodology [Journal Article]. Statistics in medicine. 2008;27(26):5525–5555.

[77] Tremblay G, Livings C, Crowe L, Kapetanakis V, Briggs A. Determination of the most appropriate method for extrapolating overall survival data from a placebo-controlled clinical trial of lenvatinib for progressive, radioiodine-refractory differentiated thyroid cancer [Journal Article]. Clinicoeconomics and Outcomes Research. 2016;8:323–333. Available from: <GotoISI>://WOS:000390477400003https://www.dovepress.com/getfile.php?fileID=31154.

[78] Negrín MA, Nam J, Briggs AH. Bayesian solutions for handling uncertainty in survival extrapolation [Journal Article]. Medical Decision Making. 2017;37(4):367–376. Available from: https://journals.sagepub.com/doi/full/10.1177/0272989X16650669?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[79] Jackson C, Stevens J, Ren S, Latimer N, Bojke L, Manca A, et al. Extrapolating survival from randomized trials using external data: a review of methods [Journal Article]. Medical Decision Making. 2017;37(4):377–390. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5424081/pdf/10.1177_0272989X16639900.pdf.

[80] Lousdal ML, Kristiansen IS, Moller B, Stovring H. Predicting Mean Survival Time from Reported Median Survival Time for Cancer Patients [Journal Article]. Med Decis Making. 2017;37(4):391–402. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27353826https://journals.sagepub.com/doi/full/10.1177/0272989X16655341?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[81] Hoogenveen RT, Boshuizen HC, Engelfriet PM, van Baal PHM. You Only Die Once: Accounting for Multi-Attributable Mortality Risks in Multi-Disease Models for Health-Economic Analyses [Journal Article]. Med Decis Making. 2017;37(4):403–414. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27405746https://journals.sagepub.com/doi/full/10.1177/0272989X16658661?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[82] Meacock R, Sutton M, Kristensen SR, Harrison M. Using Survival Analysis to Improve Estimates of Life Year Gains in Policy Evaluations [Journal Article]. Med Decis Making. 2017;37(4):415–426. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27311650https://journals.sagepub.com/doi/full/10.1177/0272989X16654444?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

[83] Williams C, Lewsey JD, Mackay DF, Briggs AH. Estimation of Survival Probabilities for Use in Cost-effectiveness Analyses: A Comparison of a Multi-state Modeling Survival Analysis Approach with Partitioned Survival and Markov Decision-Analytic Modeling [Journal Article]. Med Decis Making. 2017;37(4):427–439. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27698003https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5424853/pdf/10.1177_0272989X16670617.pdf.

[84] Williams C, Lewsey JD, Briggs AH, Mackay DF. Cost-effectiveness Analysis in R Using a Multi-state Modeling Survival Analysis Framework: A Tutorial [Journal Article]. Med Decis Making. 2017;37(4):340–352. Available

from: https://www.ncbi.nlm.nih.gov/pubmed/27281337https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5424858/pdf/10.1177_0272989X16651869.pdf.

[85] Abraira V, Muriel A, Emparanza JI, Pijoan JI, Royuela A, Plana MN, et al. Reporting quality of survival analyses in medical journals still needs improvement. A minimal requirements proposal [Journal Article]. Journal of clinical epidemiology. 2013;66(12):1340–1346.e5. Available from: https://www.jclinepi.com/article/S0895-4356(13)00250-3/fulltext.

[86] Batson S, Greenall G, Hudson P. Review of the Reporting of Survival Analyses within Randomised Controlled Trials and the Implications for Meta-Analysis [Journal Article]. PLoS One. 2016;11(5):e0154870. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27149107https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858202/pdf/pone.0154870.pdf.

[87] Boucquemont J, Heinze G, Jager KJ, Oberbauer R, Leffondre K. Regression methods for investigating risk factors of chronic kidney disease outcomes: the state of the art [Journal Article]. BMC nephrology. 2014;15(1):45.

[88] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis–an introduction to concepts and methods [Journal Article]. Br J Cancer. 2003;89(3):431–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12888808https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394368/pdf/89-6601119a.pdf.

[89] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis Part III: multivariate data analysis – choosing a model and assessing its adequacy and fit [Journal Article]. Br J Cancer. 2003;89(4):605–11. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12915864https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2376927/pdf/89-6601120a.pdf.

[90] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses [Journal Article]. Br J Cancer. 2003;89(2):232–8. Available

from: https://www.ncbi.nlm.nih.gov/pubmed/12865907https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/pdf/89-6601118a.pdf.

[91] Clark T, Bradburn M, Love S, Altman D. Survival analysis part IV: further concepts and methods in survival analysis [Journal Article]. The British Journal of Cancer. 2003;89(5):781. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394469/pdf/89-6601117a.pdf.

[92] Fleming TR, Lin D. Survival analysis in clinical trials: past developments and future directions [Journal Article]. Biometrics. 2000;56(4):971–983. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2000.0971.x?sid=nlm%3Apubmed.

[93] Oakes D, Peterson DR. Survival methods: additional topics [Journal Article]. Circulation. 2008;117(22):2949–55. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18519862.

[94] Guyot P, Welton NJ, Ouwens MJ, Ades AE. Survival time outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness [Journal Article]. Value Health. 2011;14(5):640–6. Available from: https://www.valueinhealthjournal.com/article/S1098-3015(11)00120-3/pdf.

[95] Ishak KJ, Kreif N, Benedict A, Muszbek N. Overview of parametric survival analysis for health-economic applications [Journal Article]. Pharmacoeconomics. 2013;31(8):663–675. Available from: https://link.springer.com/article/10.1007%2Fs40273-013-0064-3.

[96] Jacobs P, Golmohammadi K, Longobardi T. Lifetime costs for medical services: a methodological review [Journal Article]. Int J Technol Assess Health Care. 2003;19(2):278–86.

[97] Argyropoulos C, Unruh ML. Analysis of Time to Event Outcomes in Randomized Controlled Trials by Generalized Additive Models [Journal Article]. Plos One. 2015;10(4). Available from: <GotoISI>://WOS:000353332000045https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0123784&type=printable.

[98] Crowther MJ, Lambert PC. A general framework for parametric survival analysis [Journal Article]. Statistics in Medicine. 2014;33(30):5280–5297. Available from: <GotoISI>://WOS:000346055000006https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6300.

[99] Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis [Journal Article]. Statistics in Medicine. 2014;33(22):3844–3858. Available from: <GotoISI>://WOS:000341824600005https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6191.

[100] Jackson CH. flexsurv: A Platform for Parametric Survival Modeling in R [Journal Article]. Journal of Statistical Software. 2016;70(8):1–33. Available from: <GotoISI>://WOS:000384912000001.

[101] Titman AC. Estimation of time-shift models with application to survival calibration in health technology assessment [Journal Article]. Statistics in Medicine. 2016;35(20):3645–3656. Available from: <GotoISI>://WOS:000380728800014https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6951.

[102] Cox C, Chu H, Schneider MF, Muñoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution [Journal Article]. Statistics in medicine. 2007;26(23):4352–4374.

[103] Putter H, Fiocco M, Geskus R. Tutorial in biostatistics: competing risks and multi-state models [Journal Article]. Statistics in medicine. 2007;26(11):2389–2430.

[104] Perme MP, Andersen PK. Checking hazard regression models using pseudo-observations [Journal Article]. Statistics in medicine. 2008;27(25):5309–5328. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2749183/pdf/nihms112706.pdf.

[105] Desquilbet L, Mariotti F. Dose-response analyses using restricted cubic spline functions in public health research [Journal Article]. Statistics in medicine. 2010;29(9):1037–1057.

[106] Pullenayegum EM, Cook RJ. The analysis of treatment effects for recurring episodic conditions [Journal Article]. Stat Med. 2010;29(14):1539–58. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20535764https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3882.

[107] Govindarajulu US, Lin H, Lunetta KL, D'Agostino R. Frailty models: applications to biomedical and genetic studies [Journal Article]. Statistics in medicine. 2011;30(22):2754–2764.

[108] Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments [Journal Article]. Stat Med. 2014;33(7):1242–58. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24122911https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285179/pdf/sim0033-1242.pdf.

[109] Prenen L, Braekers R, Duchateau L. Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2017;79(2):483–505. Available from: http://dx.doi.org/10.1111/rssb.12174https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12174.

[110] Barrett J, Diggle P, Henderson R, Taylor-Robinson D. Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2015;77(1):131–148.

[111] Yang F, Small DS. Using post-outcome measurement information in censoring-by-death problems [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2016;78(1):299–318.

[112] Li R, Peng L. Quantile regression adjusting for dependent censoring from semicompeting risks [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2015;77(1):107–130. Available from: http://dx.doi.org/10.1111/rssb.12063https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12063.

[113] Chen YH. Semiparametric marginal regression analysis for dependent competing risks under an assumed copula [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010;72(2):235–251. Available from: http://dx.doi.org/10.1111/j.1467-9868.2009.00734.xhttps://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2009.00734.x.

[114] Martinussen T, Vansteelandt S, Gerster M, Hjelmborg JvB. Estimation of direct effects for survival data by using the Aalen additive hazards model [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73(5):773–788.

[115] Paddy Farrington C, Unkel S, Anaya-Izquierdo K. The relative frailty variance and shared frailty models [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012;74(4):673–696. Available from: http://dx.doi.org/10.1111/j.1467-9868.2011.01021.xhttps://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.01021.x.

[116] Cheng Y, Fine JP. Cumulative incidence association models for bivariate competing risks data [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012;74(2):183–202. Available from: http://dx.doi.org/10.1111/j.1467-9868.2011.01012.xhttps://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.01012.x.

[117] Yu W, Chen K, Sobel ME, Ying Z. Semiparametric transformation models for causal inference in time-to-event studies with all-or-nothing compliance [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2015;77(2):397–415. Available from: http://dx.doi.org/10.1111/rssb.12072https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12072.

[118] Mao L, Lin DY. Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2017;79(2):573–587. Available from: http://dx.doi.org/10.1111/rssb.12177https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12177.

[119] Rabe-Hesketh S. Multilevel and longitudinal modeling using Stata. Vol. 2, Categorical responses, counts, and survival. 3rd ed. College Station, Tex.: College Station, Tex. : Stata Press Publication, 2012; 2012.

[120] Campbell MJ. Medical statistics [electronic resource] : a textbook for the health sciences. 4th ed. Chichester, England Hoboken, NJ: Chichester, England Hoboken, NJ : Wiley, ©2007; 2007.

[121] Wienke A. Correlated frailty models in survival analysis. London: London : Chapman and Hall, 2010; 2010.

[122] Crowder MJ. Multivariate survival analysis and competing risks. Boca Raton: Boca Raton : CRC Press, 2012; 2012.

[123] Mills M. Introducing survival and event history analysis. Sage Publications; 2011.

[124] Royston P, Lambert PC. Flexible parametric survival analysis using Stata: beyond the Cox model. United States of America: Stata Press; 2011.

[125] Cleves MA. An introduction to survival analysis using Stata. 3rd ed. College Station, Tex.: College Station, Tex. : Stata Press, 2010; 2010.

[126] Barton B. Medical statistics : a guide to SPSS, data analysis and critical appraisal. Second edition. ed. Chichester : Wiley-Blackwell, 2014; 2014.

[127] Tutz G, Schmid M. Modeling discrete time-to-event data. Springer; 2016.

[128] Gamerman D. Dynamic Bayesian models for survival data [Journal Article]. Applied Statistics. 1991;p. 63–79.

[129] Davis D. An analysis of some failure data [Journal Article]. Journal of the American Statistical Association. 1952;47(258):113–150.

[130] Peng Y, Dear KB, Denham J. A generalized F mixture model for cure rate estimation [Journal Article]. Statistics in medicine. 1998;17(8):813–830.

[131] Cox C. The generalized F distribution: an umbrella for parametric survival analysis [Journal Article]. Statistics in medicine. 2008;27(21):4301–4312.

[132] Kearns B, Ara R, Young T, Relton C. Association between body mass index and health-related quality of life, and the impact of self-reported long-term conditions–cross-sectional study from the south Yorkshire cohort dataset [Journal Article]. BMC Public Health. 2013;13(1):1009.

[133] Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. vol. 777. John Wiley and Sons; 2008.

[134] Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, et al. Modelling the Survival Outcomes of Immuno-Oncology Drugs in Economic Evaluations: A Systematic Approach to Data Analysis and Extrapolation [Journal Article]. PharmacoEconomics. 2017;p. 1–14.

[135] Wood SN. Generalized additive models: an introduction with R (Second edition). CRC press; 2017.

[136] Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects [Journal Article]. Statistics in medicine. 2002;21(15):2175–2197.

[137] Crowther MJ, Lambert P. STMIX: Stata module to fit two-component parametric mixture survival models; 2016. Available from: http://EconPapers.repec.org/RePEc:boc:bocode:s457339.

[138] Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis [Journal Article]. Biostatistics. 2006;8(3):576–594.

[139] Rodriguez G. Statistical issues in the analysis of reproductive histories using hazard models [Journal Article]. Ann N Y Acad Sci. 1994;709:266–79. Available from: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1994.tb30415.x?sid=nlm%3Apubmed.

[140] Jakobsen LH, Andersson TML, Biccler JL, El-Galaly TC, Bøgsted M. Estimating the loss of lifetime function using flexible parametric relative survival models [Journal Article]. BMC medical research methodology. 2019;19(1):23. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350283/pdf/12874_2019_Article_661.pdf.

[141] Cleves M. An introduction to survival analysis using Stata. Stata Press; 2008.

[142] Jackson CH. Multi-state models for panel data: the msm package for R [Journal Article]. Journal of Statistical Software. 2011;38(8):1–29.

[143] Fahrmeir L. Dynamic modelling and penalized likelihood estimation for discrete time survival data [Journal Article]. Biometrika. 1994;81(2):317–330.

[144] He J, McGee DL, Niu X. Application of the Bayesian dynamic survival model in medicine [Journal Article]. Stat Med. 2010;29(3):347–60. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20014356https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3795.

[145] Nelder JA, Wedderburn RWM. Generalized Linear Models [Journal Article]. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1972;135(3):15.

[146] Dobson AJ, Barnett A. An introduction to generalized linear models. CRC press; 2008.

[147] Breslow N. Covariance analysis of censored survival data [Journal Article]. Biometrics. 1974;p. 89–99.

[148] Aitkin M, Clayton D. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM [Journal Article]. Applied Statistics. 1980;p. 156–163.

[149] Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve [Journal Article]. Journal of the American statistical Association. 1988;83(402):414–425.

[150] West M, Harrison PJ, Mignon HS. Dynamic Generalized Linear Models and Bayesian Forecasting [Journal Article]. Journal of the American Statistical Association. 1985;80(389):11.

[151] McKenzie E, Gardner Jr ES. Damped trend exponential smoothing: a modelling viewpoint [Journal Article]. International Journal of Forecasting. 2010;26(4):661–665.

[152] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2014.

[153] Biller C. Discrete duration models combining dynamic and random effects [Journal Article]. Lifetime data analysis. 2000;6(4):375–390.

[154] Bastos LS, Gamerman D. Dynamic survival models with spatial frailty [Journal Article]. Lifetime Data Anal. 2006;12(4):441–60. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17031498https://link.springer.com/article/10.1007%2Fs10985-006-9020-2.

[155] Wagner H. Bayesian estimation and stochastic model specification search for dynamic survival models [Journal Article]. Statistics and Computing. 2009;21(2):231–246. Available from: https://link.springer.com/article/10.1007%2Fs11222-009-9164-5.

[156] Wilson KJ, Farrow M. Bayes linear kinematics in a dynamic Bayesian survival model [Journal Article]. arXiv preprint arXiv:14112497. 2014;.

[157] Omori Y, Johnson RA. Efficient Semiparametric Bayesian Estimation of Multivariate Discrete Proportional Hazards Model with Random Effects [Journal Article]. Communications in Statistics - Theory and Methods. 2008;38(1):29–41. Available from: https://www.tandfonline.com/doi/full/10.1080/03610920802155478.

[158] Fahrmeir L, Wagenpfeil S. Smoothing hazard functions and time-varying effects in discrete duration and competing risks models [Journal Article]. Journal of the American Statistical Association. 1996;91(436):1584–1594.

[159] Sinha D, Dey DK. Semiparametric Bayesian Analysis of Survival Data [Journal Article]. Journal of the American Statistical Association. 1997;92(439):1195–1212. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474077.

[160] Klein JP, Van Houwelingen HC, Ibrahim JG, Scheike TH. Handbook of survival analysis (Chapter 2). CRC Press; 2016.

[161] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk [Journal Article]. Statistics in medicine. 2000;19(17-18):2555–2567.

[162] Kozumi H. Bayesian analysis of discrete survival data with a hidden Markov chain [Journal Article]. Biometrics. 2000;56(4):1002–1006. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2000.01002.x?sid=nlm%3Apubmed`.

[163] Sinha D. Posterior likelihood methods for multivariate survival data [Journal Article]. Biometrics. 1998;p. 1463–1474.

[164] De Waal C. A Bayesian model for estimating the failure rate for different groups [Journal Article]. South African Statistical Journal. 1995;29(2):131–154.

[165] Wong MC, Lam KF, Lo EC. Analysis of multilevel grouped survival data with time-varying regression coefficients [Journal Article]. Stat Med. 2011;30(3):250–9. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/21213342https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4094`.

[166] Yin G, Nieto-Barajas LE. Bayesian cure rate model accommodating multiplicative and additive covariates [Journal Article]. Statistics and Its Interface. 2009;2(4):513–521.

[167] dos Santos TR, Gamerman D, da Conceição Franco G. Reliability analysis via non-Gaussian state-space models [Journal Article]. IEEE Transactions on Reliability. 2017;66(2):309–318.

[168] Fahrmeir L, Knorr-Held L. Dynamic Discrete-time Duration Models: Estimation via Markov Chain Monte Carlo [Journal Article]. Sociological Methodology. 1997;27(1):417–452.

[169] Gamerman D. Bayes estimation of the piece-wise exponential distribution [Journal Article]. IEEE Transactions on Reliability. 1994;43(1):128–131.

[170] Hennerfeind A, Brezger A, Fahrmeir L. Geoadditive Survival Models [Journal Article]. Journal of the American Statistical Association. 2006;101(475):1065–1075. Available from: `https://www.tandfonline.com/doi/abs/10.1198/016214506000000348`.

[171] Kim S, Chen MH, Dey DK, Gamerman D. Bayesian dynamic models for survival data with a cure fraction [Journal Article]. Lifetime Data Anal. 2007;13(1):17–35. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/17136621https://link.springer.com/article/10.1007%2Fs10985-006-9028-7`.

[172] Gamerman D. A dynamic approach to the statistical analysis of point processes [Journal Article]. Biometrika. 1992;79(1):39–50.

[173] Ristic B, Arulampalam S, Gordon N. Beyond the Kalman filter: Particle filters for tracking applications. Artech house; 2003.

[174] Triantafyllopoulos K. Inference of Dynamic Generalized Linear Models: On-Line Computation and Appraisal [Journal Article]. International Statistical Review. 2009;77(3):430–450.

[175] Gamerman D, West M. An application of dynamic survival models in unemployment studies [Journal Article]. The Statistician. 1987;p. 269–274.

[176] Engel B, Keen A, Clayton D, Londford N, Kuba J, Firth D, et al. Hierarchical generalized linear models-Discussion [Journal Article]. Journal of the Royal Statistical Society Series B-Methodological. 1996;58(4):656–678.

[177] Gelman A, Carlin JB, Stern HS, Dunson DB. Bayesian data analysis. vol. 2. 3rd ed.; 2014.

[178] Kedem B, Fokianos K. Regression models for time series analysis. vol. 488. John Wiley and Sons; 2005.

[179] He J. Bayesian dynamic survival models for longitudinal aging data. The Florida State University; 2007.

[180] Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS book: A practical introduction to Bayesian analysis. CRC press; 2012.

[181] De Waal C. A Bayesian model for estimating the failure rate for different groups [Journal Article]. South African Statistical Journal. 1995;29(2):131–154.

[182] dos Santos TR, Gamerman D, da Conceição Franco G. Reliability analysis via non-Gaussian state-space models [Journal Article]. IEEE Transactions on Reliability. 2017;66(2):309–318.

[183] Pohar M, Stare J. Relative survival analysis in R [Journal Article]. Computer methods and programs in biomedicine. 2006;81(3):272–278. Available from: `https://www.sciencedirect.com/science/article/abs/pii/S0169260706000228?via%3Dihub`.

[184] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival [Journal Article]. Statistics in medicine. 2004;23(1):51–64.

[185] Robert CP, Casella G. Introducing monte carlo methods with r. vol. 18. Springer; 2010.

[186] Monnahan CC, Thorson JT, Branch TA. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo [Journal Article]. Methods in Ecology and Evolution. 2017;8(3):339–348.

[187] Neal RM. MCMC using Hamiltonian dynamics [Journal Article]. Handbook of markov chain monte carlo. 2011;2(11):2.

[188] Girolami M, Calderhead B. Riemann manifold langevin and hamiltonian monte carlo methods [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73(2):123–214.

[189] Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo [Journal Article]. Journal of Machine Learning Research. 2014;15(1):1593–1623.

[190] Gelman A, Lee D, Guo J. Stan: A probabilistic programming language for Bayesian inference and optimization [Journal Article]. Journal of Educational and Behavioral Statistics. 2015;40(5):530–543.

[191] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language [Journal Article]. Journal of statistical software. 2017;76(1).

[192] Team SD. RStan: the R interface to Stan; 2018. Available from: http://mc-stan.org/.

[193] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper) [Journal Article]. Bayesian analysis. 2006;1(3):515–534.

[194] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods [Journal Article]. Stat Med. 2019;48(11):2074–2102. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8086.

[195] Bland JM, Altman DG. Statistics Notes: The use of transformation when comparing two means [Journal Article]. Bmj. 1996;312(7039):1153. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2350653/pdf/bmj00540-0047.pdf.

[196] [Electronic Article]; 2019. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190308008V.

[197] Ross JS, Waldstreicher J, Bamford S, Berlin JA, Childers K, Desai NR, et al. Overview and experience of the YODA Project with clinical trial data sharing after 5 years [Journal Article]. Scientific data. 2018;5(1):1–14.

[198] De Bono JS, Logothetis CJ, Molina A, Fizazi K, North S, Chu L, et al. Abiraterone and increased survival in metastatic prostate cancer [Journal Article]. New England Journal of Medicine. 2011;364(21):1995–2005. Available from: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1014618?articleTools=true.

[199] Fizazi K, Scher HI, Molina A, Logothetis CJ, Chi KN, Jones RJ, et al. Abiraterone acetate for treatment of metastatic castration-resistant prostate cancer: final overall survival analysis of the COU-AA-301 randomised, double-blind, placebo-controlled phase 3 study [Journal Article]. The lancet oncology. 2012;13(10):983–992.

[200] [Web Page]; 2019. Available from: www.mortality.org.

[201] Burnham KP, Anderson D. Model selection and multi-model inference [Journal Article]. A Pratical informatio-theoric approch Sringer. 2003;1229.

[202] Kass RE, Raftery AE. Bayes factors [Journal Article]. Journal of the american statistical association. 1995;90(430):773–795.

[203] Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study [Journal Article]. Journal of Statistical Computation and Simulation. 2015;85(4):777–793.

[204] Gneiting T. Making and evaluating point forecasts [Journal Article]. Journal of the American Statistical Association. 2011;106(494):746–762.

[205] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building [Journal Article]. Statistics in medicine. 2007;26(30):5512–5528.

[206] Fokianos K, Tjøstheim D. Log-linear Poisson autoregression [Journal Article]. Journal of Multivariate Analysis. 2011;102(3):563–578.

[207] Magee L. Nonlocal behavior in polynomial regressions [Journal Article]. The American Statistician. 1998;52(1):20–22.

[208] Ouwens MJ, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations [Journal Article]. PharmacoEconomics. 2019;37(9):1129–1138. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6830404/pdf/40273_2019_Article_806.pdf.

[209] Bullement A, Latimer NR, Gorrod HB. Survival extrapolation in cancer immunotherapy: a validation-based case study [Journal Article]. Value in Health. 2019;22(3):276–283. Available from: https://www.valueinhealthjournal.com/article/S1098-3015(18)36202-8/pdf.

[210] Grant TS, Burns D, Kiff C, Lee D. A Case Study Examining the Usefulness of Cure Modelling for the Prediction of Survival Based on Data Maturity [Journal Article]. PharmacoEconomics. 2019;p. 1–11.

[211] [Web Page]; 2018. Available from: https://www.nice.org.uk/guidance/ta554/chapter/3-Committee-discussion#survival-modelling-in-the-economic-model.

[212] [Audiovisual Material]; 2016. Available from: http://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/previousReleases.

[213] Ltd BMSP. Single technology appraisal: Nivolumab for previously treated locally advanced or metastatic squamous non small cell lung cancer [ID811]: Company evidence submission; 2015. Available from: https://www.nice.org.uk/guidance/TA483.

[214] Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WE, Poddubskaya E, et al. Nivolumab versus docetaxel in advanced squamous-cell non–small-cell lung cancer [Journal Article]. New England Journal of Medicine. 2015;373(2):123–135. Available from: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1504627?articleTools=true.

[215] Reviews L, Group I. Single technology appraisal: Nivolumab for previously treated locally advanced or metastatic squamous non small cell lung cancer [ID811]: Evidence Review Group report; 2015. Available from: https://www.nice.org.uk/guidance/TA483.

[216] Ltd BMSP. BMS response to the Appraisal Consultation Document (ACD) for nivolumab for previously treated locally advanced or metastatic squamous non-small-cell lung cancer (NSCLC); 2016. Available from: https://www.nice.org.uk/guidance/TA483.

[217] Reviews L, Group I. Bristol-Myers Squibb (BMS) response to the Appraisal Consultation Document (ACD) for nivolumab for previously treated locally advanced or metastatic squamous non-small cell lung cancer (NSCLC) Evidence Review Group (ERG) commentary on issues raised; 2016. Available from: https://www.nice.org.uk/guidance/TA483.

[218] [Web Page]; 2019. Available from: https://markummitchell.github.io/engauge-digitizer/.

[219] Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves [Journal Article]. BMC medical research methodology. 2012;12(1):9.

[220] Baio G. survHE: Survival analysis for health economic evaluation and cost-effectiveness modelling [Journal Article]. Journal of Statistical Software. 2020;Accepted for publication. Available from: `https://gianlucabaio.netlify.com/publication/baio-2019/`.

[221] Woods B, Sideris E, Palmer S, Latimer N, Soares M. In: NICE Decision Support Unit Technical Support Documents. London: National Institute for Health and Care Excellence (NICE) unless otherwise stated. All rights reserved.; 2017. .

[222] Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. Oup Oxford; 2006.

[223] Unit NDS. Single technology appraisal: Nivolumab for previously treated locally advanced or metastatic squamous non small cell lung cancer [ID811]: Comments on the ongoing appraisals of nivolumab for squamous and non-squamous non-small cell lung cancer; 2016. Available from: `https://www.nice.org.uk/guidance/TA483`.

[224] Alarid-Escudero F, Kuntz KM. Potential Bias Associated with Modeling the Effectiveness of Healthcare Interventions in Reducing Mortality Using an Overall Hazard Ratio [Journal Article]. PharmacoEconomics. 2020;38(3):285–296.

[225] Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies [Journal Article]. Cancer Epidemiology. 2012;36(1):16–21.

[226] Ades A, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis [Journal Article]. Pharmacoeconomics. 2006;24(1):1–19.

[227] Bullement A, Willis A, Amin A, Schlichting M, Hatswell AJ, Bharmal M. Evaluation of survival extrapolation in immuno-oncology using multiple pre-planned data cuts: learnings to aid in model selection [Journal Article]. BMC Medical Research Methodology. 2020;20(1):1–14.

[228] Gelfand A, Gilks W, Richardson S, Spiegelhalter D. Markov chain Monte Carlo in practice [Journal Article]. Boca Rotan, FL: Chapman and Hall. 1996;p. 145–161.

[229] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines [Journal Article]. The journal of chemical physics. 1953;21(6):1087–1092.

[230] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications [Journal Article]. Biometrika. 1970;57:13.

[231] Wilkinson DJ. Stochastic modelling for systems biology. Chapman and Hall/CRC; 2006.

[232] Girolami M, Calderhead B. Riemann manifold langevin and hamiltonian monte carlo methods [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73(2):123–214.

[233] West M, Harrison J. Bayesian forecasting and dynamic models. Springer Science and Business Media; 2006.

[234] Helske J. KFAS: Kalman filter and smoothers for exponential family state space models [Journal Article]. R package version. 2014;1:4–1.

[235] Morgan BJ. Elements of simulation. vol. 4. CRC Press; 1984.

[236] Gardner Jr ES, McKenzie E. Forecasting trends in time series [Journal Article]. Management Science. 1985;31(10):1237–1246.

[237] Makridakis S, Hibon M. The M3-Competition: results, conclusions and implications [Journal Article]. International journal of forecasting. 2000;16(4):451–476.

[238] Nixon RM, Bansback N, Stevens JW, Brennan A, Madan J. Using short-term evidence to predict six-month outcomes in clinical trials of signs and symptoms in rheumatoid arthritis [Journal Article]. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry. 2009;8(2):150–162.

[239] White IR. simsum: Analyses of simulation studies including Monte Carlo error [Journal Article]. The Stata Journal. 2010;10(3):369–385.

[240] Mandel M. Simulation-based confidence intervals for functions with complicated derivatives [Journal Article]. The American Statistician. 2013;67(2):76–81.

[241] Hawkins N, Grieve R. Extrapolation of Survival Data in Cost-effectiveness Analyses: The Need for Causal Clarity [Journal Article]. Med Decis Making. 2017;37(4):337–339.

[242] StataCorp. Stata Statistical Software: Release 14. StataCorp LP; 2015.

[243] Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2016. Available from: https://www.R-project.org/.

[244] Vickers A. An Evaluation of Survival Curve Extrapolation Techniques Using Long-Term Observational Cancer Data [Journal Article]. Medical Decision Making. 2019;39(8):926–938.

[245] Strauss DJ, Vachon PJ, Shavelle RM. Estimation of future mortality rates and life expectancy in chronic medical conditions [Journal Article]. J Insur Med. 2005;37(1):20–34.

[246] Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Longitudinal data analysis. CRC Press; 2008.

[247] Fahrmeir L, Tutz G. Multivariate statistical modelling based on generalized linear models. Springer Science and Business Media; 2013.

[248] Hastie T, Tibshirani R. Generalized additive models. Wiley Online Library; 1990.

[249] Jackson CH, Thompson SG, Sharples LD. Accounting for uncertainty in health economic decision models by using model averaging [Journal Article]. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2009;172(2):383–404.

[250] [Web Page]; 2018. Available from: https://robjhyndman.com/hyndsight/forecasting-competitions/.

[251] Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response [Journal Article]. Statistics in Medicine. 2013;32(13):2262–2277.

[252] Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015.

[253] Peixoto JL. A property of well-formulated polynomial regression models [Journal Article]. The American Statistician. 1990;44(1):26–30.

[254] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis [Journal Article]. Stata Journal. 2009;9(2):265.

[255] Remontet L, Uhry Z, Bossard N, Iwaz J, Belot A, Danieli C, et al. Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: Performance of this multidimensional penalized spline approach in net survival trend analysis [Journal Article]. Statistical methods in medical research. 2018;p. 0962280218779408.

[256] Molenberghs G, Verbeke G. Models for discrete longitudinal data. Springer; 2005.

[257] Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models [Journal Article]. Journal of the American Statistical Association. 2004;99(467):673–686.

[258] Baayen RH, van Rij J, de Cat C, Wood SN. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models [Journal Article]. arXiv preprint arXiv:160102043. 2016;.

[259] Wood SN. Thin plate regression splines [Journal Article]. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003;65(1):95–114.

Appendix 1

Technical appendix

1.1 Survival data and analyses

This section introduces some of the key concepts for survival analyses and describes how they are related. Survival data measure the elapsed time until the event of interest, relative to a defined time origin, such as the time of randomisation in a clinical trial. Typically, not all individuals have experienced the event of interest at the end of follow-up. Such individuals are said to have (right-)censored event times. In these situations the true event time is not known, only that it will be greater than the observed censored time. Reasons for censoring include if the individual is event-free at the end of data collection, or if they are lost to further follow-up before experiencing the event of interest. Survival models are required for the analysis of censored data. Standard survival models are described in Section 1.1.2, key introductory concepts are provided below.

Standard approaches to analysing survival data assume that censoring is uninformative [17]. That is, conditional on any covariates that have been collected, patients with censored outcomes are no different to patients without censored outcomes. Analyses of informative censoring are beyond the scope of this thesis.

1.1.1 Key concepts

The probability density function $f(t)$: For an individual, let t denote their observed event time. This can be viewed as an observation of a random variable T which can take any non-negative number ($T \in \mathbb{R}_{\geq 0}$). The distribution of T is the probability density function. The probability that an event occurs before time t (the *cumulative density function*) is then:

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{1.1}$$

The hazard function $h(t)$: This is derived from the probability that an event occurs within the interval (t to $t + \Delta t$) conditional on the event not having occurred by (or before) time t. Dividing this conditional probability by the interval width Δt turns it into a rate (probability per unit time). The hazard function $h(t)$ is obtained as the limiting value of this rate as $\Delta t \to 0$. This function is also referred to as the *instantaneous event rate*.

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T < t + \Delta t \mid T \geq t}{\Delta t} \right\} \tag{1.2}$$

The survival function $S(t)$: This is a cumulative distribution, which represents the probability that an individual survives beyond time t:

$$S(t) = P(T \geq t) = 1 - F(t) = 1 - \int_0^t f(u)du \tag{1.3}$$

It is related to the hazard function:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\{\log S(t)\} \tag{1.4}$$

The cumulative hazard function $H(t)$: This is the integral of the hazard as defined in Equation (1.2). It may be interpreted as either the cumulative risk of an event occurring by time t or the expected number of events by time t. It may also be derived from the survivor function.

$$H(t) = \int_0^t h(u)du \tag{1.5}$$

$$= -\log S(t) \tag{1.6}$$

Equations (1.1), (1.3) and (1.2) are linked by the relationship:

$$f(t) = h(t) \times S(t) \tag{1.7}$$

Hazard estimation

Life-table (actuarial) estimates: These group the observed survival times into a series of m time intervals. For example, if survival is measured in years, the time intervals may be two years: 0 to 2 years, 2 to 4 years, 4 to 6 years, and so on. To remove ambiguity if an observed survival time is equal to the end (or start) of an interval (say $t^* = 4$), the intervals are defined as $t_{j-1} \leq t^* < t_j$, or equivalently $[t_{j-1}, t_j)$ (so a survival time of $t^* = 4$ would be assigned to the interval '4 to 6 years'). The width of the j-th interval (Δ_j) is the time between the start and end times of the interval. For the example given, $\Delta_j = 2$ for all intervals, in general Δ_j can vary for each interval. Also, for the j-th interval, let n_j be the number of individuals alive at the start of the interval, d_j be the number of events in the interval and c_j be the number of censored times within the interval. Under the actuarial assumption that both the censoring and event times occur at a constant rate within the interval (are uniformly distributed), the number at risk, hazard rate and its (asymptotic) standard error (se) during the jth time interval may be estimated as:

$$\begin{aligned} n_j^* &= n_j - c_j/2 \\ h^*(t) &= \frac{d_j}{(n_j^* - d_j/2)\Delta_j} \\ \text{se}\{h^*(t)\} &= \frac{h^*(t)\sqrt{\{1 - [h^*(t)\Delta_j/2]^2}}{\sqrt{(d_j)}} \end{aligned}$$

Kaplan-Meier type estimate: This uses the observed survival times t^*. Unlike the life-table estimate, there is no grouping of event times. It should be noted that events are only observed at discrete times and it is assumed that the underlying continuous hazard function is constant during the interval between event times. If there are d_j events and n_j individuals at risk at time $t_{(j)}$ and the time until the next observed event time $t_{(j+1)}$ is Δ_j, then:

$$\begin{aligned}\hat{h}(t) &= \frac{d_j}{n_j \Delta_j} \\ \text{se}\{\hat{h}(t)\} &= \hat{h}(t)\sqrt{\frac{n_j - d_j}{n_j d_j}}\end{aligned}$$

Likelihood estimation

Let δ_i be an event indicator which $= 1$ if t_i^* is an observed survival time and $= 0$ if it is a right-censored observation (hence the true survival time will be greater than t_i^*). The likelihood is then:

$$\ell_i = \left\{ \prod_{i:\sim\delta_i=1} f_i(t_i^*) \prod_{i:\sim\delta_i=0} S_i(t_i^*) \right\} \tag{1.8}$$

Where $\prod$ denotes multiplication. For individuals with an observed event their pdf contributes to the likelihood. Individuals with censored times contribute their cumulative survivor function (as their pdf is not fully observed but it is known that they survived up to time t_i). The log-likelihood contribution for the ith patient is:

$$\begin{aligned}\log \ell_i &= \log \left\{ f(t_i^*)^{\delta_i} S(t_i^*)^{1-\delta_i} \right\} \\ &= d_i \log\{f(t_i^*)\} + (1 - \delta_i) \log\{S(t_i^*)\}\end{aligned} \tag{1.9}$$

Substituting Equations (1.7) and (1.3) into the above gives:

$$\begin{aligned}\log \ell_i &= \log \left\{ h(t_i^*)^{\delta_i} S(t_i^*) \right\} \\ &= \delta_i \log\{h(t_i^*)\} - \int_0^{t_i^*} h(u) du\end{aligned} \tag{1.10}$$

1.1.2 Standard distributions (models)

For the following distributions both the survivor distribution $S(t)$ and the pdf $f(t)$ are presented. If a simple expression for the hazard function $h(t)$ is available, this is also provided. Otherwise it is noted that this may be derived from Equation (1.7). With regards to terminology, terms 'distribution' and 'model' are used interchangeably. The below specifications are primarily based on Collett, 2015 [17]

Exponential (λ)

$$
\begin{aligned}
S(t) &= e^{-\lambda t} \\
f(t) &= \lambda e^{-\lambda t} \\
h(t) &= \lambda
\end{aligned}
$$

The notable feature of the exponential distribution is that the hazard is a constant value at all times.

Weibull (λ, γ)

$$
\begin{aligned}
S(t) &= e^{-\lambda t^{\gamma}} \\
f(t) &= \lambda \gamma t^{\gamma-1} e^{-\lambda t^{\gamma}} \\
h(t) &= \lambda \gamma t^{\gamma-1}
\end{aligned}
$$

When $\gamma = 1$, the Weibull is the same as the exponential distribution. Otherwise, the hazard function is either monotone increasing or monotone decreasing over time (that is, there are no turning points in the hazard function). This property may be easier to identify when considering the logarithm (log) of the hazard (as the logarithm of a monotone function is itself monotone):

$$\begin{aligned}\log[h(t)] &= \log(\lambda\gamma t^{\gamma-1}) \\ &= \log((\lambda) + \log(\gamma) + \log(t^{\gamma-1}) \\ &= c + (\gamma - 1)\log(t)\end{aligned}$$

Where $c = \log((\lambda) + \log(\gamma)$ is a constant, so the log-hazard function is a linear function of log-time.

Gompertz (λ, θ)

$$\begin{aligned}S(t) &= \exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t})\right\} \\ f(t) &= \lambda e^{\theta t} \exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t})\right\} \\ h(t) &= \lambda e^{\theta t} = \exp(\alpha + \theta t)\end{aligned}$$

The Gompertz distribution has similar properties to the Weibull distribution. For the Gompertz, the log-hazard function is a linear function of time:

$$\begin{aligned}\log[h(t)] &= \log(\lambda e^{\theta t}) \\ &= \log(\lambda) + \log(e^{\theta t})) \\ &= c + \theta t\end{aligned}$$

When $\theta = 0$ the Gompertz is the same as the exponential distribution. As with the Weibull, the hazard function is otherwise monotone increasing or monotone decreasing.

Log-logistic (θ, κ)

$$\begin{aligned} S(t) &= (1 + e^{\theta} t^{\kappa})^{-1} \\ f(t) &= \frac{e^{\theta} \kappa t^{\kappa - 1}}{(1 + e^{\theta} t^{\kappa})^2} \\ h(t) &= \frac{e^{\theta} \kappa t^{\kappa - 1}}{1 + e^{\theta} t^{\kappa}} \end{aligned}$$

The log-logistic is more flexible than the Weibull and Gompertz distributions, as it can model a unimodal hazard (that is, a hazard function with a turning-point), which occurs for $\kappa > 1$. For $\kappa \leq 1$ the log-logistic models a monotonic decreasing hazard. The log-logistic distribution does not include the exponential distribution as a special case.

Lognormal (μ, θ)

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\theta}\right)$$

where $\Phi(\dot{})$ is the standard normal distribution function:

$$\begin{aligned} \Phi(z) &= \frac{1}{\sqrt{(2\pi)}} \int_{-\inf}^{z} \exp(-u^2/2) du \\ f(t) &= \frac{1}{\theta\sqrt{(2\pi)}} t^{-1} \exp\{-(\log t - \mu)^2 / 2\theta^2\} \end{aligned}$$

It can be seen that the hazard function involves an integral, which makes interpretation difficult. It has been noted however that, for any given dataset, the fitted log-logistic distribution is very similar to the fitted lognormal distribution with the lognormal θ parameter multiplied by 1.82 [102, 17]. As with the log-logistic distribution, the lognormal distribution allows for unimodal hazard functions but does not simplify to an exponential.

Gamma ($\lambda\beta$)[102, 17]

$$S(t) = 1 - \Gamma[\beta t; \lambda]$$

$$\Gamma[\beta t; \lambda] = \frac{1}{\Gamma(\lambda)} \int_0^{\beta t} u^{\lambda-1} e^{-u} du$$

$$f(t) = \frac{\beta^{\lambda} t^{\lambda-1} e^{-\beta t}}{\Gamma(\lambda)}$$

As with the Weibull and Gompertz distributions, the gamma distribution is always monotonic and has the exponential distribution as a special case when $\lambda = 1$.

Generalised Gamma (λ, θ, β) [102]

$$\begin{aligned} S(t) &= 1 - \Gamma[\lambda^{-2} \exp(\lambda[\log(t) - \beta]/\theta); \lambda^{-2}] \\ (\text{if } \lambda > 0) &= 1 - \Gamma[\lambda^{-2}(e^{-\beta} t)^{\lambda/\theta}; \lambda^{-2}] \\ (\text{if } \lambda < 0) &= \Gamma[\lambda^{-2}(e^{-\beta} t)^{\lambda/\theta}; \lambda^{-2}] \end{aligned}$$

$$f(t) = \frac{|\lambda|}{\theta t \Gamma(\lambda^{-2})} [\lambda^{-2}(e^{-\beta} t)^{\lambda/\theta}]\lambda^{-2} \exp([-\lambda^{-2}(e^{-\beta} t)^{\lambda/\theta}])$$

This three-parameter model includes the exponential, Weibull, gamma, and lognormal disributions as special cases (with $\lambda = \theta = 1, \lambda = 1, \lambda = \theta$, and $\lambda \to 0$ respectively). The hazard function may be a constant, monotone increasing or decreasing, 'bath-tub' shaped (initially decreasing before a turning-point and then increasing), or arc-shaped.

Generalised F (β, σ, m_1, m_2) [131]

$$f(t) = \frac{e^{-\beta m_1/\sigma} t^{(m_1/\sigma)-1} (m_1/m_2)^{m_1}}{\sigma B(m_1, m_2)[1 + (m_1/m_2)(e^{-\beta} t)^{1/\sigma}]^{m_1+m_2}}$$

where $B(m_1, m_2)$ is the beta function evaluated at m_1, m_2.

The equation for the survival function is not straightforward but it may be obtained via equation (1.3).

The generalised F includes both the log-logistic and the generalised gamma distributions as special cases ($m_1 = m_2 = 1$ for the former and either $m_1 \to \infty$ or $m_2 \to \infty$ for the latter). As such, it can take any of the shapes covered by these two distributions.

1.2 Parameter estimation for dynamic models

Three main estimation methods are currently used for DSMs: Linear Bayes, posterior mode estimation, and MCMC. The technical details of these three methods are provided here.

1.2.1 Markov chain Monte Carlo

MCMC algorithms estimate the posterior distribution of model parameters: $p(\theta|y)$ where θ represents the model parameters and y represents the available data. Inferences typically require integrating the posterior distribution. For example, to estimate the mean of the model parameters requires evaluating $E[\theta|y] = \int_\theta \theta \; p(\theta|y) \; d\theta$ [180]
MCMC algorithms may be thought of as a general method for estimating integrals. A particular strength of MCMC algorithms is their flexibility as they may applied to almost any complex problem [228]. As their name suggests, MCMC algorithms have two main components: the use of *Markov chains* and *Monte Carlo* simulation.

Integration via Monte Carlo simulation relies on drawing a sample from the target distribution $(X_1, \ldots, X_n)$ and using the empirical average to approximate the integral:

$$E_f[h(x)] = \int h(x)f(x)dx = \frac{1}{N}\sum_{i=1}^{N} h(x_i) \tag{1.11}$$

A Markov chain has the property that the current distribution of X at time t only depends on the value of X at the previous time-points and not on any of the preceding values (states). The Markov chains used in MCMC algorithms have certain additional properties which ensure that it will have a stationary probability distribution (See Chapter 6 of Robert and Casella for more details [185]) . Hence, if $X^{(t)} \sim f$ then $X^{(t+1)} \sim f$ where the superscript denotes the iteration number. Over time the values from such a Markov chain will produce samples from the target density. These samples may then be used to perform Monte Carlo integration - the combination of these two approaches is formalised in the framework of MCMC algorithms. These are iterative algorithms: as the number of iterations increases, the approximating distribution created by the Markov chain converges towards the target (posterior) distribution. This convergence will occur independently of the starting values chosen. As such MCMC algorithms represent a powerful and flexible method for analysing statistical problems. However, a key limitation with these algorithms is that they only provide information on the areas of f that they have explored, with no indication of how much of f remains unexplored. In other words "you've only seen where you've been" [185] (p238). As such, it is important that the Markov chain explores all of the target density: the properties of Markov chains used in MCMC algorithms ensure this asymptotically but do not provide any guarantee that this will happen for a finite sequence.

The Metropolis-Hastings algorithm

An initial MCMC algorithm was proposed by Metropolis *et al* in 1953 and generalised by Hastings in 1970 [229, 230]. The resulting Metropolis-Hastings (M-H) MCMC algorithm is:

Given $x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t \text{ with probability } p(x^{(t)}, Y_t), \\ x^{(t)} \text{ with probability } 1 - p(x^{(t)}, Y_t), \end{cases}$$

where

$$p(x,y) = \min\left[\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right]$$

and $q()$ is a candidate (proposal) distribution, that may be simulated from [185]. The value $p(x, y)$ is the acceptance probability. Averaging this across iterations provides the *acceptance rate*.

The basic idea underlying the M-H algorithm is as follows. For a Markov chain in a given state, a proposal state is simulated from $q()$. If this state is more likely to come from the target distribution f, then the Markov chain moves to the proposed state. Otherwise it remains in its current state. The M-H algorithm includes the original algorithm by Metropolis *et al* as a special case, for which $q()$ is a random walk centred on the current value of $x^{(t)}$ (and the ratio involving $q()$ cancels) . As such, the M-H algorithm explores the local neighbourhood of the current value of the Markov chain $x^{(t)}$. Another special-case of the M-H algorithm occurs if each component of θ is updated once per iteration, which is known as *Gibbs sampling* [228].

An advantage of an MCMC algorithm with a random walk is that it can be used in virtually any situation. However, use of a random walk is also inefficient: it does not take advantage of the characteristics of the target function (such as local gradients), so will return to areas of the target that have already been explored. Due to the symmetric nature of the random walk, about half of all iterations will re-visit already explored parts of the target density. As with all MCMC algorithms, the M-H algorithm also requires tuning, which can be difficult. In particular, it is typically unclear what the ideal value of the acceptance rate would be. High acceptance rates suggest both that the Markov chain is exploring different parts of the target density (which is good) and also that the chain has not yet converged (which is not good). Conversely, low acceptance rates suggest that the chain has converged but provide no guarantee that it has explored enough of the target density (Wilkinson, Chapter 9 [231]). There are also computational implications: lower acceptance rates will produce chains with higher correlation, which leads to a loss of efficiency. Higher acceptance rates will reduce correlations but typically require smaller transitions of the Markov chain. As such, it will take longer to explore the target density [232].

1.2.2 Filtering methods

Filtering algorithms provide a method for sequentially updating posterior distributions. For example, with a filtering algorithm, estimates of model coefficients at time t_i use all the evidence *up to and including* time t_i. This is in contrast to MCMC methods, which are smoothing methods as an estimate at t_i uses all of the available data [173]. After obtaining filtered estimates, it is possible to use all of the data to smooth these, as described here.

If a dynamic generalised linear model is used with the Normal distribution, the optimal filtering algorithm is that described by Rudolf Kalman in 1960 [173], so this approach is typically referred to as *Kalman filtering.* This algorithm is optimal as that it provides an exact description of the posterior (which is Normally distributed, so may be fully described by its mean and covariance).

It also provides minimum least-squares estimates of one-step ahead prediction [178]. DSMs have non-Gaussian outcomes so require alternative filtering approaches such as posterior mode estimation and linear Bayes. Both share similarities with the Kalman filter, so this approach is described first.

Kalman Filter

The Kalman filter uses the following derived variables:

One-step ahead forecast errors: $e_{t_i} = y_{t_i} - \hat{y}_{t_i|t_{i-1}} = y_{t_i} - zx'_{t_i}\hat{\beta}_{t_i|t_{i-1}}$

Kalman gain: $k_{t_i} = P_{t_i|t_{i-1}} z_{t_i} / q_{t_i|t_{i-1}}$, where $q_{t_i|t_{i-1}} = z'_{t_i} P_{t_i|t_{i-1}} z_t t_i + \sigma^2$.

Further, it is assumed that the distribution of the states is normal, with mean $\hat{\beta}$ and variance-covariance matrix P.
It has four main steps:

1. Initialization:

$$\beta_0 \sim N(\hat{\beta}_{0|0}, P_{0|0})$$

2. Filter prediction (for $t_i = 1, \ldots, T$):

$$\begin{aligned} \hat{\beta}_{t_i|t_{i-1}} &= F\hat{\beta}_{t_{i-1}|t_{i-1}} \\ P_{t_i|t_{i-1}} &= FP_{t_{i-1}|t_{i-1}}F' + Z_t \end{aligned}$$

3. Filter correction (for $t_i = 1, \ldots, T$):

$$\begin{aligned} \hat{\beta}_{t_i|t_i} &= \hat{\beta}_{t_i|t_{i-1}} + k_{t-1}e_{t-1} \\ P_{t_i|t_i} &= P_{t_i|t_{i-1}} - q_{t_i|t_{i-1}} k_{t-1} k'_{t-1}. \end{aligned}$$

4. Filter smoothing (for $t_i = T, T-1, \ldots, 1$):

$$\begin{aligned}
\hat{\beta}_{t_{i-1}|T} &= \hat{\beta}_{t_{i-1}|t_{i-1}} + B_t(\hat{\beta}_{t_i|T} - \hat{\beta}_{t_i|t_{i-1}}) \\
P_{t_{i-1}|T} &= P_{t_{i-1}|t_{i-1}} + B_{t_i}(P_{t_i|T} - P_{t_i|t_{i-1}})B_t' \\
B_{t_i} &= P_{t_{i-1}|t_{i-1}} F_{t_i}' P_{t_i|t_{i-1}}^{-1}
\end{aligned}$$

The Kalman gain is used during filter correction. It controls both the degree to which state estimates are effected by the one-step ahead errors and also the amount of decrease in the state variances. There are four stages in a filtering algorithm. The first is initial of the algorithm: in a Bayesian context this involves specifying prior values. The second is to generate one-step ahead predictions (extrapolations, or forecasts). The process of forecasting the (unobserved) future creates additional uncertainty so the variance of the states increases. These forecasts are then corrected in the third step, using data for the new time-period. This correction leads to filtered estimates, for which the variance decreases due to using more data. The final step generates smoothed estimates. It is implemented after all of the prediction and correction steps have been estimated.

The Kalman gain may also be interpreted as proportion of the total variance attributed to the states. For given values of σ^2, the larger the uncertainty in the parameter estimates, the more they are ‘corrected’ by the error value (e_{t_i}). For given values of $P_{t_i|ti-1}$, larger (smaller) values of σ^2 implies smaller (larger) precision in the observations (were precision is defined as $1/\sigma^2$) so a smaller (larger) correction occurs. As such, the Kalman gain may be interpreted as controlling the relative trade-off between faith with previous model estimates ($k_{t_i} \to 0$) and faith with the observed data ($k_{t_i} \to 1$).

Linear Bayes

This is implemented in a similar way to the Kalman filter but it does not make any distributional assumptions [128]. As such, it is applicable in non-Gaussian settings. However, to make inferences about either the states (β values) or the model estimates of the outcome, it is necessary to assume that these follow a distribution.

1. Initialization. Note that the distribution of the βs is no longer specified.:

$$\beta_0 \sim (\hat{\beta}_{0|0}, P_{0|0})$$

2. Prediction is the same as for the Kalman filter.

3. Correction is based on conjugate analysis and differs from the Kalman filter. It is defined at the patient-level (denoted by the subscript j). For a Poisson model, where the outcome is the hazard λ_{t_i} for times $t_i = 1, \ldots, T$:

$$\begin{aligned} \hat{\beta}_{t_ij|t_i} &= \hat{\beta}_{t_ij|t_{i-1}} + \frac{P_{t_ij|t_{i-1}} z_j'}{q_{t_ij}} \log\left(\frac{1 + q_{t_ij}\delta_{t_ij}}{1 + q_{t_ij}(\omega_{t_i})\exp(-z_j\hat{\beta}_{t_ij|t_{i-1}})}\right) \\ P_{t_ij|t_i} &= P_{t_ij|t_{i-1}} - \frac{\delta_{t_ij}}{1 + q_{t_ij}}(P_{t_ij|t_{i-1}} z_j')(P_{t_ij|t_{i-1}} z_j')' \end{aligned}$$

   where $q_{t_ij} = z_j P_{t_ij|t_{i-1}} z_j'$, δ_{t_i} is an indicator of if the individual died in the interval (=1, else = 0). Values are updated across patients with $\hat{\beta}_{t_i,j+1|t_{i-1}} = \hat{\beta}_{t_ij|t_i}, P_{t_i,j+1|t_{i-1}} = P_{t_ij|t_i}$

4. Smoothing is the same as for the Kalman filter.

As with the Kalman filter, the correction step leads to a decrease in the variance of the states; this decrease only occurs for patients who experienced an event during the interval. A difference to the Kalman filter is that with linear Bayes, parameter estimates are updated one patient at a time (as opposed to one-time-period at a time). Resulting estimates can depend on the order of updating (if patients with $\delta_{t_ij} = 1$ are updated first or second), which is an undesirable feature of linear Bayes [156]. It also assumes that Z_{t_i} is known (in contrast, most of the other estimation methods allow for unknown Z_{t_i}.

The linear Bayes approach also shares many similarities with posterior mode estimation, which is described next.

Posterior mode estimation

This approach obtains a linear approximation to the response function used in the DSM. This approximation is based on local linearisation, using a Taylor series expansion [173, 233]. This approach is also known as a Laplace approximation [234]. The description here is implemented using the extended Kalman filter, so may be viewed as a filtering method. In general, let $X' = g(X)$ be the transformation of the random variable X by the response function $g(\cdot)$. Denote the probability density function of X by $f_X(\cdot)$ and denote the probability density function of X' by $f'_X(\cdot)$. The two are related by the equation:

$$f'_X(X') = f_X[g^{-1}(X')] \left| \frac{d}{dX'} g^{-1}(X') \right| \tag{1.12}$$

the term $\left|\frac{d}{dX'} g^{-1}(X')\right|$ is known as the 'Jacobian' (Chapter 2 of [235]).
As an example, let $X' = \exp(X)$, hence $g(\cdot) = \exp(\cdot)$, $g^{-1}(\cdot) = \log(\cdot)$ and $\frac{d}{dX'} \log(\cdot) = 1/(\cdot)$. Summary measures for the transformed variable X' (such as the mean and variance) are obtained from $f'_X(X')$. This demonstrates that a Jacobian adjustment may be used to calculate outcomes of interest when non-linear transformations (in this context, response functions) are used. One of their uses is in posterior mode estimation

In general, posterior mode estimation may be thought of as approximating the (non-linear) DSM by a dynamic linear model, for which the Kalman filter may be applied. The approximation is based around the posterior mode of the DSM. The algorithm for posterior mode estimation is as follows:

1. Initialization:

$$\beta_0 \sim N(\hat{\beta}_{0|0}, P_{0|0})$$

2. Filter prediction (for $t = 1, \ldots, T$):

$$\hat{\beta}_{t_i|t_{i-1}} = F\hat{\beta}_{t_{i-1}|t_{i-1}}$$

$$P_{t_i|t_{i-1}} = FP_{t_{i-1}|t_{i-1}}F' + Z_{t_i}$$

3. Filter correction (for $t_i = 1, \ldots, T$):

$$\begin{aligned}\hat{\beta}_{t_i|t_i} &= \hat{\beta}_{t_i|t_{i-1}} + P_{t_i|t_i} u_t \\ P_{t_i|t_i} &= 1/(1/P_{t_i|t_{i-1}} + U_{t_i})\end{aligned}$$

where

$$\begin{aligned}u_{t_i} &= \sum z_{t_ij} \frac{d_{t_ij}}{\hat{\lambda}_{t_ij|t_{i-1}}(1-\hat{\lambda}_{t_ij|t_{i-1}})}(\lambda_{t_ij} - \hat{\lambda}_{t_ij|t_{i-1}}) \\ U_{t_i} &= \sum z_{t_ij} z'_{t_ij} \frac{d_{t_ij}}{\hat{\lambda}_{t_ij|t_{i-1}}(1-\hat{\lambda}_{t_ij|t_{i-1}})}\end{aligned}$$

where d_{tj} is the Jacobian of the response function evaluated for individual j at time t (using any relevant covariates x_t), and summation is across at-risk individuals.

Note that when the response function is the logistic function, the Jacobian will evaluate to $\hat{\lambda}_{tj|t-1}(1 - \hat{\lambda}_{tj|t-1})$, so the fraction in both u_t and U_t will cancel.

4. Filter smoothing (for $t_i = T, T-1, \ldots, 1$):

$$\begin{aligned}\hat{\beta}_{t-1|T} &= \hat{\beta}_{t_{i-1}|t_{i-1}} + B_{t_i}(\hat{\beta}_{t_i|T} - \hat{\beta}_{t_i|t_{i-1}}) \\ P_{t-1|T} &= P_{t_{i-1}|t_{i-1}} + B_{t_i}(P_{t_i|T} - P_{t_i|t_{i-1}})B'_{t_i} \\ B_{t_i} &= P_{t_{i-1}|t_{i-1}} F' P^{-1}_{t_i|t_{i-1}}\end{aligned}$$

The initialisation, prediction and smoothing steps are the same as for the Kalman filter (so the prediction and smoothing steps of posterior mode estimation are also the same as those for linear Bayes).

The filter correction step differs from that used in the Kalman filter but it has the same interpretation, with $P_{t_i|t_i}$ the equivalent of the kalman gain k_{t_i} and the error being quantified by u_{t_i} (which incorporates an adjustment for non-linearity).

The main differences between the three approaches (Kalman filter, linear Bayes and posterior mode estimation) are the implementation of the filter correction step. For the Kalman gain the correction step adjusts the filtered state estimate by the product of the kalman gain k_{t_i} and the one-step ahead forecast error. The Kalman gain varies between zero (when the variance of the states is zero) and one (as the variance of the states $\rightarrow \infty$) and may be thought of as measuring the degree of faith that is placed in the previous state estimate ($k_{t_i} = 0$ implies perfect faith with the state estimates, $k_{t_i} = 1$ implies no faith). Both the linear Bayes and posterior mode estimation approaches share this interpretation albeit with different implementation due to the non-linearity of the DSM. A key difference between the linear Bayes and posterior mode estimation is (when the same response function and distribution for the outcomes are used) how residuals are defined. For posterior mode estimation residuals are based on the difference between observed and modelled estimates of the hazard $\lambda_{t_ij} - \hat{\lambda}_{t_ij|t_{i-1}}$. For linear Bayes the residuals are based on the approximate difference between prior and posterior estimates of $z_j\hat{\beta}_{t_ij|t_{i-1}}$.

A final difference of note is that with posterior mode estimation (implemented with the extended Kalman filter), the filter-correction for the $\hat{\beta}_{t_i}$ uses the filtered value of the variance $P_{t_i|t_i}$. In contrast, both the Kalman filter and linear Bayes approaches use the predicted value $P_{t_i|t_{i-1}}$

1.3 Model specification for dynamic survival models

To improve clarity, the descriptions in this section omit the subscript i from time; that is $t, t-1$ here correspond to t_i, t_{i-1}.

1.3.1 Local level model with a random walk

Model specification for the linear predictor is:

$$\begin{aligned} g(y_t) = \eta_t &= \beta_t \\ \beta_t &= \beta_{t-1} + \zeta_t \\ &= \beta_{t-2} + \zeta_t + \zeta_{t-1} \\ &\dots \\ &= \beta_0 + \sum_{i=1}^{t} \zeta_i \end{aligned}$$

with $\zeta_t \sim \text{Normal}(0, Z_t)$. It is often assumed that $Z_t = Z \ \forall \ t$.

Extrapolations: The extrapolated value h time steps into the future, given observed data up to time T is $\hat{\eta}_{T+h|T} = \beta_T$.

Variance: The variance at time t is $\sum_{i=1}^{t} Z_i$ (or $t \times Z$ if the ζ_t are constant).

Limitations: As $t \to \infty$ then the variance of the $\beta_t \to \infty$, which is unlikely to be realistic. The assumption of a constant extrapolated hazard is also unlikely to be realistic.

1.3.2 Local trend model with a random walk

These may be defined as:

$$\begin{aligned}
\eta_t &= \beta_{1t} \\
\beta_{1t} &= \beta_{1,t-1} + \beta_{2,t-1} + \zeta_{1,t} \\
\beta_{2t} &= \beta_{2,t-1} + \zeta_{2,t}
\end{aligned}$$

Hence:

$$\begin{aligned}
\beta_{1t} &= \beta_{1,t-1} + \beta_{2,t-1} + \zeta_{1,t} \\
&= \beta_{1,t-2} + 2\beta_{2,t-2} + \zeta_{1,t} + \zeta_{1,t-1} + \zeta_{2,t-1} \\
&= \beta_{1,t-3} + 3\beta_{2,t-3} + \zeta_{1,t} + \zeta_{1,t-1} + \zeta_{1,t-2} + \zeta_{2,t-1} + 2\zeta_{2,t-2} \\
&= \ldots \\
&= \beta_{1,0} + t\beta_{2,0} + \sum_{i=1}^{t} \zeta_{1i} + \sum_{j=1}^{t-1} j\zeta_{2,t-j}
\end{aligned}$$

where the first two terms are a linear function and the last two terms are random error.

Extrapolations: $\hat{\eta}_{T+h|T} = \beta_{1,T} + h\beta_{2,T}$.

Variance: $\sum_{i=0}^{t} Z_t + \sum_{j=1}^{t-1} (jZ_{2,t-j})$.

If we assume that the Z_t are constant over time we get $t \times Z_1 + (\sum j = 1^{t-1})^2 \times Z_2$.

Limitations: As with the local-level random walk model, the variance of the local-trend random walk model will $\to \infty$ as $t \to \infty$ (this holds even if we were to remove the local-level component). In addition, there is empircal evidence (albeit not from the field of survival analysis), that assuming a persistent trend during the extrapolated period generally leads to worse forecasts than if that trend is 'dampened' (reduced) over-time [236, 237]. However, a response function is applied to the extrapolations, which will have a similar effect to dampening (current empirical evidence is for when there is no transformation of the trends), so it is unclear if this empirical evidence generalises to both the field of survival analysis and situations where the trend is transformed via a response function.

A final limtation is the potential lack of identifiability due to the flexibility of this model. In particular, a local-level can visually look like a local-trend, so it may not be possible to accurately estimate both.

1.3.3 Local trend model with an autoregressive process

A model which allows seperate AR processes for the level and trend is given by:

$$\begin{aligned}
E(y_t) &= \beta_{1t} \\
\beta_{1t} &= \phi_1\beta_{1,t-1} + \phi_2\beta_{2,t-1} + \zeta_{1,t} \\
\beta_{2t} &= \phi_2\beta_{2,t-1} + \zeta_{2,t}
\end{aligned}$$

hence:

$$\begin{aligned}
\beta_{1t} &= \phi_1^2\beta_{1,t-2} + \phi_2^2\beta_{2,t-2} + \phi_2\phi_1\beta_{2,t-2} + \zeta_{1,t} + \phi_1\zeta_{1,t-1} + \phi_2\zeta_{2,t-1} \\
&= \phi_1^3\beta_{1,t-3} + \phi_2^3\beta_{2,t-3} + \phi_2^2\phi_1\beta_{2,t-3} + \phi_2\phi_1^2\beta_{2,t-3} \\
&\qquad + \zeta_{1,t} + \phi_1\zeta_{1,t-1} + \phi_1^2\zeta_{1,t-2} + \phi_2\zeta_{2,t-1} + \phi_2^2\zeta_{2,t-2} + \phi_2\phi_1\zeta_{2,t-2} \\
&= \phi_1^4\beta_{1,t-4} + \phi_2^4\beta_{2,t-4} + \phi_2^3\phi_1\beta_{2,t-4} + \phi_2^2\phi_1^2\beta_{2,t-4} + \phi_2\phi_1^3\beta_{2,t-4} \\
&\qquad + \zeta_{1,t} + \phi_1\zeta_{1,t-1} + \phi_1^2\zeta_{1,t-2} + \phi_1^3\zeta_{1,t-3} \\
&\qquad + \phi_2\zeta_{2,t-1} + \phi_2^2\zeta_{2,t-2} + \phi_2\phi_1\zeta_{2,t-2} + \phi_2^3\zeta_{2,t-3} + \phi_2^2\phi_1\zeta_{2,t-3} + \phi_2\phi_1^2\zeta_{2,t-3} \\
&= \ldots \\
&= \phi^t\beta_{1,0} + \sum_{i=1}^{t}\left(\phi_2^j\phi_1^{t-j}\beta_{2,0}\right) + \sum_{j=0}^{t-1}\left(\phi^j\zeta_{t-j}\right) + \sum_{\ell=1}^{t-1}\left[\sum_{k=1}^{t-\ell}\left(\phi_2^k\phi_1^{(t-k-1)}\zeta_{2,\ell}\right)\right].
\end{aligned}$$

A special case is if only the trend has an AR process with Z_t constant. The model simplifies to:

$$E(y_t) = \beta_{1,0} + \sum_{i=1}^{t} \left(\phi_2^j \beta_{2,0}\right) + \sum_{j=0}^{t-1} (\zeta_{t-j}) + \sum_{\ell=1}^{t-1} \left[\sum_{k=1}^{t-\ell} \left(\phi_2^k \zeta_{2,\ell}\right)\right]. \tag{1.13}$$

Below equations for extrapolations and variance relate to this special model.

Extrapolations: $\hat{\eta}_{T+h|T} = \beta_{1,0} + \sum_{i=1}^{h} (\phi_2^i)\beta_{2,T}$.

The last term is a geometric progression, so as $t \to \infty$ its sum shall become $\phi_2\beta_{2,T}/(1-\phi)$. For a large enough time-horizon, extrapolations shall converge to a constant value (that is, the trend is dampened over time until it becomes zero). Note that this will also hold if the local-level has an AR process.

Variance: $\sum_{j=0}^{t-1} (Z_{t-j}) + \sum_{\ell=1}^{t-1} \left[\sum_{k=1}^{t-\ell} \left(\phi_2^k Z_{2,\ell}\right)\right]$.

It has already been demonstrated that the first term has a finite sum. Consider the inner-part of the second term: $\sum_{k=1}^{t-\ell} \left(\phi_2^k Z_{2,\ell}\right)$. Again appealing to the properties of a geometric progression, this has the sum:

$$Z\frac{1-\phi^{t-\ell-1}}{1-\phi} - Z\frac{1-\phi}{1-\phi} = Z\phi\frac{1-\phi^{t-\ell-2}}{1-\phi} \tag{1.14}$$

which has a finite sum as $t \to \infty$ if $|\phi| \leq 1$. There are a finite sum of these in the second term of the variance calculation so the variance remains finite.

Limitations: This model has potential issues of identifiability (see discussion for local-trend model with a random walk).

1.4 Estimating the Monte Carlo standard error when the estimand varies with observations

As motivation, consider a simulation study where the generated data are survival data. For this each simulated dataset is T observations of hazard values over time $(\lambda_1, \ldots, \lambda_T)$. The aim of the study is to compare how well different models estimate the true hazard values $\theta_i,\ i \in \{1, \ldots, T\}$. Hence the true value of the estimand is the same for all simulated datasets, but it varies with time (observations) within a simulated dataset.

Let $\hat{\theta}_i$ be the model estimate of θ_i. Then mean-squared error (MSE) and its variance may be estimated as:

$$\text{MSE}_i = \mu_i = \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \left(\hat{\theta}_{i,j} - \theta_i\right)^2$$

$$\text{Var}(\mu_i) = \sigma_i^2 = \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \sqrt{\frac{\sum_{j=1}^{n_{\text{sim}}} \left[(\hat{\theta}_{i,j} - \theta_i) - \mu_i\right]^2}{n_{\text{sim}}(n_{\text{sim}} - 1)}}$$

Note that the subscripts i, j denote time-points and simulated datasets, respectively (the estimate is also weighted by sample size as this varies over time, but this detail is omitted for clarity here).

To aid comparison across different models, there is interest in summarising MSE over time-periods:

$$f(x) = \frac{1}{T} \sum_{i=1}^{T} \mu_i \tag{1.15}$$

Calculating $E[f(x)]$ is straightforward, calculating $\text{Var}[f(x)]$ is not. In some instances the delta method may be employed to estimate $\text{Var}[f(x)]$ using Taylor-series expansions [238]. This approach is used by White to calculate $\text{Var}[f(x)]$ for some performance measures [239], but it is unclear if this approach may be used for MSE.

A simulation-based approach has been suggested as an alternative to the delta method, and has been shown to provide similar results without the need to calculate derivatives [240]. One approach to simulating $\text{Var}[f(x)]$ would be to sample $\tilde{\mu}_i \sim N(\mu_i, \sigma_i^2)$, use this to get a distribtion for $f(x)$ and derive the variance from this (assuming that the μ_i, and hence $\tilde{\mu}_i$ are independent). This approach is used for the bias as this is an unbounded performance measure. For MSE, which is constrained to be positive, an alternative distribution needs to be used instead of the Normal. An advantage of using the gamma distribution is that the mean and variance (on the scale of the data) may be used to specify the parameters: gamma(shape $= \mu_i^2/\sigma_i^2$, scale $= \sigma_i^2/\mu_i$). Hence the gamma is used for the MSE in this thesis.

Appendix 2

Further results appendix

2.1 Methodological mapping review search details

2.1.1 Bibliographic databases

Of the database reviews, the first review was designed to identify examples when methods have been used to extrapolate survival data. The search was performed in two databases; PubMed and Web of Science (WoS). For PubMed, the following search terms were used:

1. survival analyses[MeSH Terms]
2. actuarial analysis[MeSH Terms]
3. 1 or 2
4. extrapolat*[Title/Abstract]
5. forecast*[Title/Abstract]
6. 4 or 5
7. 3 and 6

For WoS, searches were restricted to the WoS Core Collection, and the following search terms were used:

1. survival[TOPIC]

2. time-to-event[TOPIC]
3. failure time[TOPIC]
4. 1 or 2 or 3
5. extrapolat*[Title/Abstract]
6. forecast*[Title/Abstract]
7. 5 or 6
8. 4 and 7

During the initial searches, an alternative search term of "expectation of life" was identified [71], so a further search for this key-word was performed in PubMed.

The second review was designed to identify review articles on the analysis of survival data. Based on the results of the first review, it was decided to focus on just PubMed, as the use of WoS provided little additional benefit. The second review had two components. The first component included any review papers. A broad statistical search term was used to keep the number of identified studies manageable. The following search terms were used:

1. survival analyses[MeSH Major Topic]
2. review[Title/Abstract]
3. review[Publication Type])
4. 2 or 3
5. 1 and 4

The second component searched for reviews specifically in the context of HTA, as it was believed that these would be of the most relevance to this work. As it was more focused, it used more detailed search terms. Note that due to the use of more statistical search terms, this search would not be a subset of the previously described search.

1. analysis, cost benefit[MeSH Terms]
2. models, economic[MeSH Terms]

3. decision support techniques[MeSH Terms]
4. "health services research/methods" [MeSH Terms]
5. "Technology Assessment, Biomedical/methods"[MeSH Terms]
6. 1 or 2 or 3 or 4 or 5
7. Survival Analysis[MeSH Terms]
8. longitudinal studies[MeSH Terms]
9. models, proportional hazards[MeSH Terms]
10. forecasting[MeSH Terms]
11. 7 or 8 or 9 or 10
12. Review[Title/Abstract]
13. Review[Publication Type]
14. 12 or 13
15. 6 and 11 and 14.

Hand searching key journals

Hand-searching was performed for two key journals. The first was the JRSS B. This journal is restricted to papers that describe methodological developments in statistics, making it suitable for hand-searching. To keep the searching to an achievable level, searching went back to the start of 2010 inclusive. The second journal was Statistics in Medicine. Searching was restricted to "tutorials in biostatistics". These are designed to provide an overview of key issues in the statistical design and analysis of data. As there was a relatively small number of tutorials, searching went back to the start of 2000 inclusive.

In addition, a Medical Decision Making special issue on "Methods for Extrapolating Survival in Cost-Effectiveness Analyses" was published in May 2017 [241]; the articles from this issue were also included in the review.

Grey literature

Two types of grey literature were considered. These were academic textbooks, and methods available within computer software. The search of textbooks was restricted to those available from the University of Sheffield. A list of textbooks was obtained using ‘StarPlus’; the University’s online library catalogue. Textbooks were identified by searching for the keyword “survival analysis”, with a pragmatic restriction to textbooks published within the last 10 years. If multiple editions of a book were available, the most recent edition was used to define the date of publication. Two different software packages were reviewed: STATA [242] and R [243]. These were chosen due to familiarity with the packages. In addition to the methods available in the base package, the methods available via user-contributed packages were also considered. For STATA this involved a search for “survival analysis, net” using the inbuilt search dialogue box. For R the CRAN task-view was used. Task views provide a list of the different statistical methods available for a given subject, and their associated packages. The survival analysis task view is available at https://cran.r-project.org/web/views/Survival.html.

Citation searching

This used the results from the search of HTA review articles as pearls. Articles citing these pearls were identified via WoS. There was no restriction on the number of citations considered.

2.2 Literature review: incorporating external data

This section describes methods for incorporating external data within extrapolations, as identified via the literature searches of Chapter 2. Relative survival models and cure models are discussed in Chapter 3, so are not replicated here. Two of the studies described here assumed that after a certain point, the hazard may be modelled as a function of age (instead of as a function of time since diagnosis) [23, 76]. Whilst this approach does not require external data, long-term external evidence on how the hazard varies with age may be used, so this approach is included here. The same terminology as Chapter 3 is used, where $\lambda_{t_i}^P$ and $\lambda_{t_i}^E$ denote the general population hazard and the disease-specific (excess) hazard respectively (both at time t_i), with the overall hazard given by $\lambda_{t_i} = \lambda_{t_i}^P + \lambda_{t_i}^E$.

2.2.1 Methods used in technology appraisals

The earliest identified description of incorporating external data is that by Hwang and Wang [70]. This method was originally developed for the extrapolation of quality-adjusted (weighted) survivor functions but a later publication demonstrates that this method may also be applied to survival functions [65]. This method has the following steps:

1. Creating (or obtaining) the external data. This external population should have long-term data on survival (so that no extrapolation is necessary). The external population should also be as similar to the internal population as possible.
2. Using the length of follow-up of the observed data, calculate the survival ratio: $\frac{\text{observed survival}}{\text{external survival}}$.
3. Fit a linear regression model to the logit-transformation of this survival ratio, with time as the only covariate.
4. For the extrapolated period, estimates of survival are obtained by multiplying the external survivor function by the inverse logit of the model estimate of the survival ratio.

The authors suggested that for step three the analysis could exclude data from early time-points if the survival ratio is not stable for these, although the choice of what data to exclude is essentially arbitrary. Bootstrapping may be used to provide an estimate of uncertainty, but it would not incorporate uncertainty in the choice of what data to exclude. The authors justify the use of a logit transformation as it is assumed that the survival ratio will never exceed 1 (in other words, the observed survival shall never be greater than the corresponding external survival). To relax this assumption, the logarithm of the survival ratio could be used instead.

Messori and Trippoli implicitly suggest using an implicit cure model [75], recommending that extrapolation be achieved by assuming that ‘long-term survivors‘ have the same life expectancy as the general population. That is, over time the hazard function for the population of interest will converge to that of the general population. Identifying the time at which cure can be assumed may not be straightforward. The practical guide to extrapolating survival data in HTA presented by Latimer [20] stressed the importance of considering external data when assessing clinical plausibility. Four examples of incorporating external data within NICE appraisals were presented, with a recommendation that further research into this area was required. The examples applied mortality rates from external life tables for the extrapolated period, and so can be viewed as equivalent to an implicit cure model. Similarly, Meacock and colleagues note that policy evaluations typically use published data on life expectancy, and so assume that long-term survival is the same as for the general population [82].

Other methods for incorporating external data for extrapolation in a HTA context are reviewed by Jackson and colleagues [79]. In addition to the above-described methods, other approaches include assuming a systematic relationship between the observed and external hazards, such as proportional hazards, additive hazards, or proportional cause-specific hazards.

2.2.2 Bayesian methods

Demiris and Sharples describe how Bayesian methods may be used to incorporate external data, which acts as prior data [68]. This is then combined with the observed data, and the resulting posterior distributions are used to generate extrapolations. A case-study compared the use of a proportional hazards Weibull model with a variety of semi-parametric models: both proportional and additive hazards were considered, along with different weights for the prior data. For the parametric Weibull model, the external data influenced the scale parameter. For the semi-parametric models, a gamma process was used to model the baseline hazard; the parameters for this were estimated based on the external data. Model-fitting was performed in WinBugs [180]. The authors noted that more flexible parametric models could have been employed. For the semi-parametric models the extended gamma process and the Beta-Stacey process were mentioned as potential alternatives for the baseline hazard. Finally, the authors noted that for certain model specifications (such as the additive semi-parametric model), the methods struggled to adequately fit both the external and observed datasets, “Hence for long term prediction, it appears that the modeler should be particularly careful in the implementation of these methods” [68]. Consequences of not having enough data were noted to be poorer predictive ability and sensitivity to the choice of prior (even when uninformative priors are used).

The methods of Demiris and Sharples were extended by Benaglia, Jackson and Sharples to model cause-specific hazards [64]. The authors consider two separate causes: deaths due to the disease of interest, and deaths from all other diseases. It was assumed that the hazard for the disease of interest is proportional between the observed and external populations, whilst the hazard for other diseases was the same in the two populations. The hazard function was modelled using a poly-Weibull model, which is a special case of the polyhazard model, as described by Demiris, Lunn and Sharples [67].

Guyot and colleagues used a Bayesian framework to incorporate three different types of external evidence: (1) general population survival, (2) registry data on one-year conditional survival, and (3) clinical input into the likely long-term treatment effect. These three types of external evidence were incorporated by defining their impact on the likelihood function and hence including them in parameter estimation [22]. The authors noted that incorporating external evidence led to an increased reliability and plausibility of extrapolations but could also lead to numerical problems during parameter estimation. This method was evaluated in a subsequent paper, which identified both difficulties with convergence and sensitivity of the method to treatment effects, as it only produced plausible extrapolations when the hazard ratio was less than 1 plus minus 0.3 [244].

2.2.3 Approaches used in demography

Day, Reynolds and Kush describe methods for estimating the life expectancy of individuals with cerebral palsy via extrapolation [66]. They recommend the use of the 'proportional life expectancy' method. This is not described in detail, but a reference is provided to Strauss, Vachon and Shavelle [245]. These authors describe five methods for extrapolation, all of which are modelled on the hazard scale, and described here. For the following, the value $\mathcal{C}$ is used for the relative effect, the definition of which depends on the approach used.

Constant excess death rate: This method assumes that $\lambda_{t_i} - \lambda^P_{t_i} = \mathcal{C}$, where $\mathcal{C}$ is a constant. Hence the excess hazard is $\lambda^E_{t_i}$ is assumed to be constant for both the duration of the observed data and for the extrapolations.

Constant relative risk: Whereas the previous method assumed a constant additive effect, this method assumes a constant relative effect: $\mathcal{C} = \lambda_{t_i} / \lambda^P_{t_i}$

'Rating Up': This method assumes that the hazard observed for the observed population at a given age (x), is equal to the hazard observed for the external population at a higher age ($x + \mathcal{C}$). Hence $\lambda_{t_i=x} = \lambda^P_{t_i=x+\mathcal{C}}$.

This method has been suggested in the context of HTA by Bagust and Beale [19]. In this situation the hazard was a function of time since diagnosis (instead of a function of age), and the observed and external data were replaced by the control and intervention study arms.

Proportional life expectancy: This applies the relative survival methodology to life expectancies. Let $\mathcal{C} = e_{t_i}/e^P_{t_i}$, where $\mathcal{C}$ is assumed to be a constant at all times, e_{t_i} is the life expectancy for an individual aged t_i in the observed population, and $e^P_{t_i}$ is the corresponding measure for the external population. The authors note that assuming proportional life expectancy is the same as assuming that the excess death rate for a given age is inversely proportional to the remaining life expectancy at that age, hence the excess death rate increases with increasing age.

Log-linear declining relative risk: This method is motivated by the observation that age-specific mortality often follows the Gompertz Law: $\lambda_{t_i} = \exp(\gamma + kt_i)$, where γ and k are constants. Under the assumption that the hazard function for both the observed and external populations follows this law (with separate constants for each), then the following equation is derived: $\log(\lambda_{t_i}/\lambda^P_{t_i}) = \beta(\alpha - t_i)$, where β and α are constants. The value of α is referred to as the *parity age* as it is the age at which the hazards of the observed and external populations become the same. The name of this method reflects the linear model for the logarithm of the relative risks. It may be viewed as relaxing the assumption of a constant relative risk to allow it to vary with time.

The authors note that these methods are designed for chronic, non-progressive conditions. This appears to be mainly because the relative effects (the $\mathcal{C}$) are assumed to be constant at all times.

2.2.4 Modelling on both the age and timescales

Information on how the hazard of mortality varies with age may be used to complement extrapolations. There were two main approaches to this: modelling extrapolations as a function of age only, or as a function of both age and time. These two approaches are discussed in turn.

Nelson and colleagues [76] describe a method which assumes that after a certain time t_i the hazard function is stable with respect to time, and so the only variation is due to ageing. Provided that the available data cover a sufficient range of ages, this method removes the need for extrapolation: the survival experience for an individual during the unobserved follow-up period is described by their ageing. The effect of ageing on the hazard is estimated directly from the sub-set of individuals who survived up-to time t_i (hence their data will be left-truncated at time t_i). The key assumption of the method described by Nelson and colleagues is that after a certain time-point any further changes in the hazard are solely due to ageing [76]. The same assumption was applied by Jackson and colleagues, who took the time t_i to be equal to five years, based on a plot of the hazard function over time [23]. In addition, Nelson and colleagues also suggest that the value for t_i could be based on clinical opinion, or it could be set equal to the maximum follow-up time. They further note that there may be insufficient data to estimate the hazard as a function of age at upper ages, and so suggest fitting a Gompertz model, whose parameters may be informed by observing the mortality at equivalent ages in the general population. These two studies were both in a HTA context [23, 76].

Extrapolations that used age and time employed a Lee-Carter model [72, 73]. This model takes the form:

$$\log \mu_{x,t_i} = \alpha_x + \beta_x \kappa_{t_i} + \epsilon_{x,t_i} \tag{2.1}$$

where μ_{x,t_i} is the hazard for an individual aged x at time t_i. The ϵ_{x,t_i} independent and identically distributed random variables following a Normal$(0, \sigma^2)$ distribution. The remaining three terms are parameters to be estimated: α_x and β_x are age-related parameters and assumed to be fixed over time. The parameter κ_{t_i} is assumed to vary with time and is modelled as a time series. In theory any timeseries model may be used to estimate the κ_{t_i}, but in practice a random walk model with drift is typically used [73].

The Lee-Carter model is notable for not including any observable quantities in the right-hand side. It is also not identifiable from the data. To make the model identifiable, constraints are typically applied to the two terms κ_{t_i} and β_x: that the latter sums to unity and the former to zero. Under this constraint, the α_x may be interpreted as the age-specific average of the observed hazard (over time). The κ_{t_i} correspond to a latent process of the trend in the hazard over time and β_x represents the effect of age on this latent process. The method described by Nelson and colleagues [76] may be thought of as an application of the Lee-Carter model, with the product $\beta_x \kappa_{t_i}$ set to zero.

Haberman and Renshaw also describe alternative specifications for the above model [72]. These include using a negative-Binomial distribution instead of a Poisson distribution, or using the probability of death as the outcome and modelling using a Binomial distribution. Majer and colleagues demonstrated that the Lee-Carter model may be applied within a multistate model with the health-states of ‘nondisabled’, ‘disabled’ and ‘dead’ [73]. There were three possible transitions: from nondisabled to disabled, and from either alive state to dead. The resulting estimates were used to derive a measure of health expectancy. There were no examples of the Lee-Carter model being used within a HTA context.

Within HTA, when extrapolating hazards, interest is usually on the change in hazard as a function of time; either since randomisation or diagnosis. However, for the outcome of mortality, it is known that the hazard varies with age. This information may be used to enhance extrapolations; a formal statistical model for this (the Lee-Carter model) is used within the demography literature. This model uses time series methods to extrapolate the hazard as a function of time, whilst also including an interaction with the effect of age and a main effect of age. There were no examples of using this model within HTA, but a simpler assumption that the hazard may be modelled as a function of age alone has been used.

2.3 Prior distributions for the innovation variance: case-studies

2.3.1 Methods

This used the two case-studies introduced elsewhere in the thesis, which were the trials of 'nivolumab vs docetaxel for previously treated squamous non-small-cell lung cancer' and 'abiraterone acetate for metastatic castration-resistant prostate cancer' [198, 214]. For this analysis the control arms are used. Three models of increasing complexity were considered: a local level model, a local trend model with a global level, and a local trend model with a local level. The available data are patient survival times, these were converted to time intervals based on unique observed death times.

Specification of initial values was not required for the local level model. For the two local trend models (with global or local level), initial values were specified for the innovation variance alone (Nivolumab case-study) or the innovation variance and the level and trend (Abiraterone case-study). Initial values were sampled from the following distributions: U(0, 0.2) for Z, U(-10, 0) for $\beta_{1,t}$ and N(0, 0.1) for $\beta_{2,t}$. The use of multiple starting values increases confidence that when the model-fitting algorithm converges, it does so to the global maximum of the likelihood function. Without specifying these initial values, models did not converge for some prior distributions. As the true value of the innovation variance was not known, performance was assessed based on three criteria: if the models converged, the effective sample size (based on the number of iterations, scaled to adjust for correlations. Larger values are better), and visual goodness of fit to the observed data.

Results

All priors converged for the local level model ($\hat{R} \leq 1.01$), with the exception of the Uniform$(-\infty, \infty)$ for log(standard deviation) prior for the Nivolumab case-study and the Gamma(0.02, 0.1) prior for the Abiraterone case-study. For the local trend global level model a lack of convergence was noted for the Uniform$(-\infty, \infty)$ for log(standard deviation) prior for both case-studies and for prior IDs (8, 10) for the Abiraterone case-study. All models converged for the local trend local level model with the exception of the Gamma(0.01, 0.01) prior for the Abiraterone case-study.

Trace plots were in visual agreement with the $\hat{R}$ statistic for assessing convergence.

Table 2.1 compares the effective sample size from each prior distribution (averaged over the three models considered) for both case studies. To aid comparison, these are scaled to be relative to the largest sample size, and results from the simulation study are also included. Due to its consistently poor performance, the Uniform$(-\infty, \infty)$ for log(standard deviation) prior (ID 2) is excluded. Visual comparisons of the model-based mean hazard estimates with 95% posterior intervals, and the observed hazards are provided for the local trend local level model (visual results were very similar for all three models) in Figure 2.1.

Table 2.1: Effective sample size relative to the largest value for each model ($\hat{R}$) for the prior distributions.

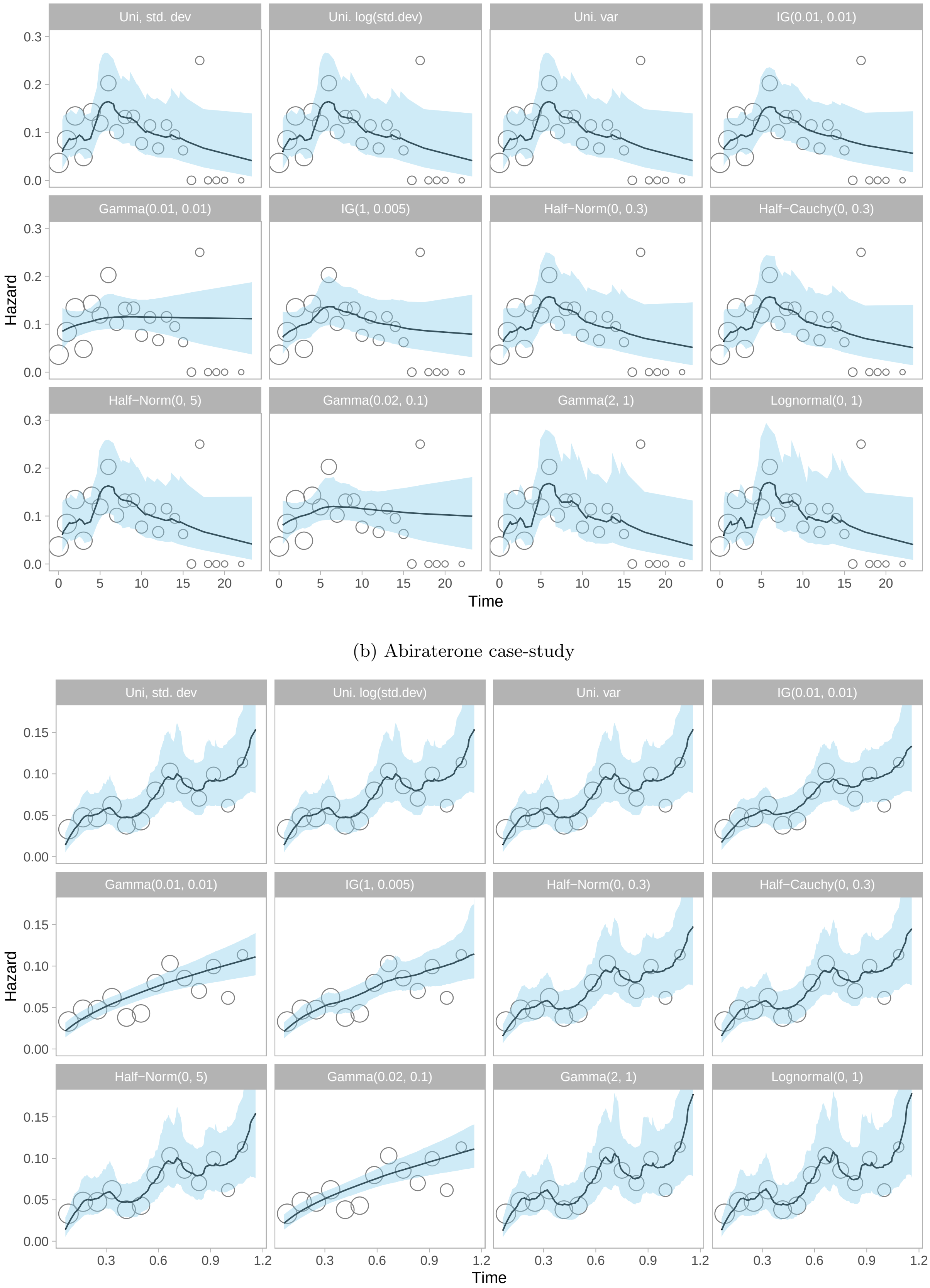
ID	**Uninformative priors**	**Nivolumab**	**Abiraterone**	**Simulation study**
1	Uniform$(0, \infty)$ std. dev.	37%	37%	62%
3	Uniform$(0, \infty)$	46%	50%	69%
4	Inv.Gamma(0.01, 0.01)	30%	55%	88%
5	Inv.Gamma(1, 0.005)	60%	46%	100%
6	Half-Cauchy(0, 0.3)	50%	59%	82%
ID	**Weakly-informative priors**			
7	Half-normal(0, 0.3)	100%	81%	67%
8	Half-normal(0, 5)	46%	53%	74%
9	Gamma(0.01, 0.01)	26%	36%	79%
10	Gamma(0.02, 0.1)	24%	12%	73%
11	Gamma(2, 1)	81%	89%	71%
12	Log-normal(0, 1)	79%	84%	71%

For both case-studies the largest effective sample sizes were observed for three weakly informative priors (IDs 7, 11, 12). However, the last two of these priors were associated with very large bias and mean squared error values in the simulation studies. In contrast, for the simulation study the largest effective sample size was observed for two uninformative priors (the two inverse gamma [IG] priors). The IG(1, 0.005) prior has the largest effective sample size for the simulation study, fourth largest for the Nivolumab case study and 8th largest for the Abiraterone case-study. This poor performance for the Abiraterone case-study is partly due a lack of convergence for one model; excluding this model, the IG(1, 0.005) prior has the 4th largest effective sample size.

For the Nivolumab case study, for all three models, prior IDs (3, 9 and 10) led to visual estimates that were too flat. All other priors led to very similar visual results. For Abiraterone, prior IDs

(9 and 10) led to visual estimates that were too flat, as did prior ID 7 for the local level model.

Figure 2.1: Comparison of model estimates and observed hazards: local trend local level model.



(a) Nivolumab case-study

(b) Abiraterone case-study

2.4 Mixture Weibull simulation study

The following graphs are displayed:

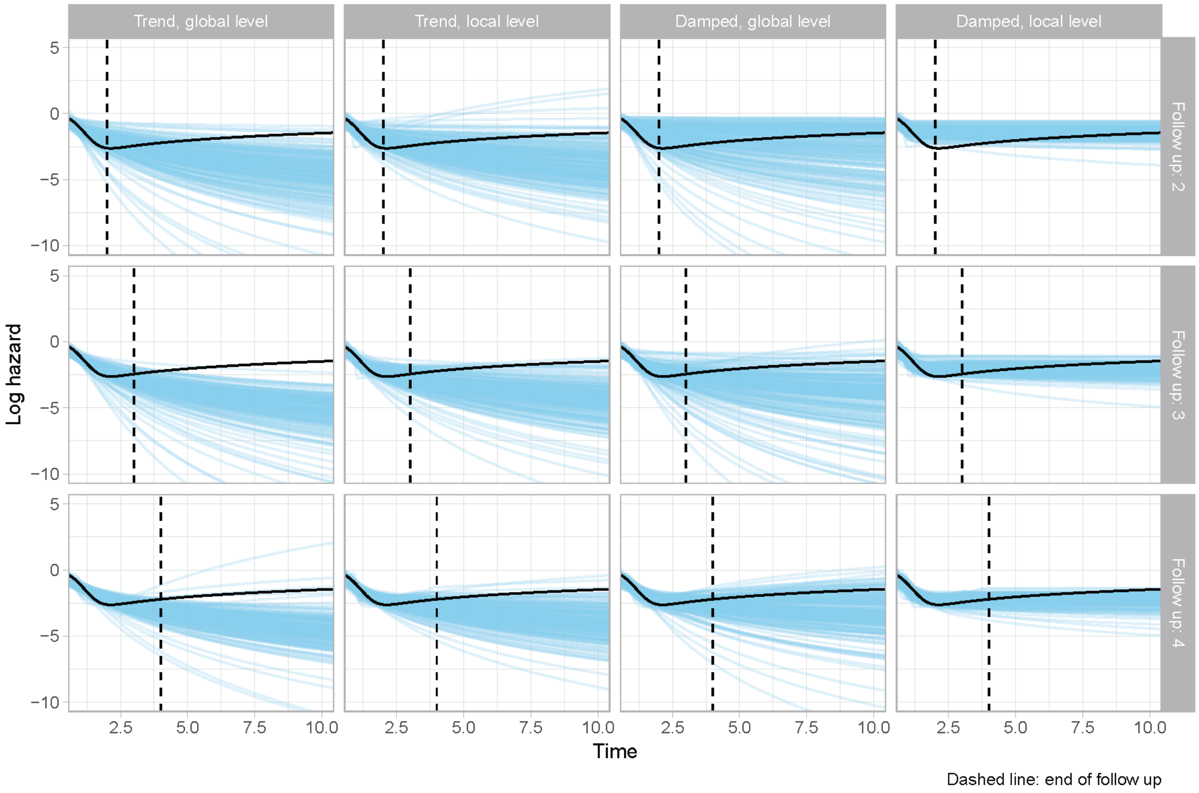
- Visual goodness of fit (comparing model estimates with the truth) are provided in Figures 2.2 and 2.3 for sample sizes of 100 and 600, respectively. This uses the same specification for the models as for the main analysis.

- Estimates of mean-squared error and bias over time are provied in Figures 2.4 and 2.5 for sample sizes of 100 and 600, respectively. This uses the same specification for the models as for the main analysis.

- The impact of using time as a covariate (instead of log-time) on visual goodness of fit is demonstrated in Figure 2.6 for a sample size of 300 and all three follow-up times (results for the other two sample sizes were very similar and are not presented). Royston-Parmar models are not included as by definition they use the logarithm of time. Both dynamic models have a global level.

Figure 2.2: Model estimates of the log-hazard (blue lines) and true values (black lines), sample size = 100.

(a) Dynamic survival models.



(b) Current and emerging practice (excluding Gompertz).

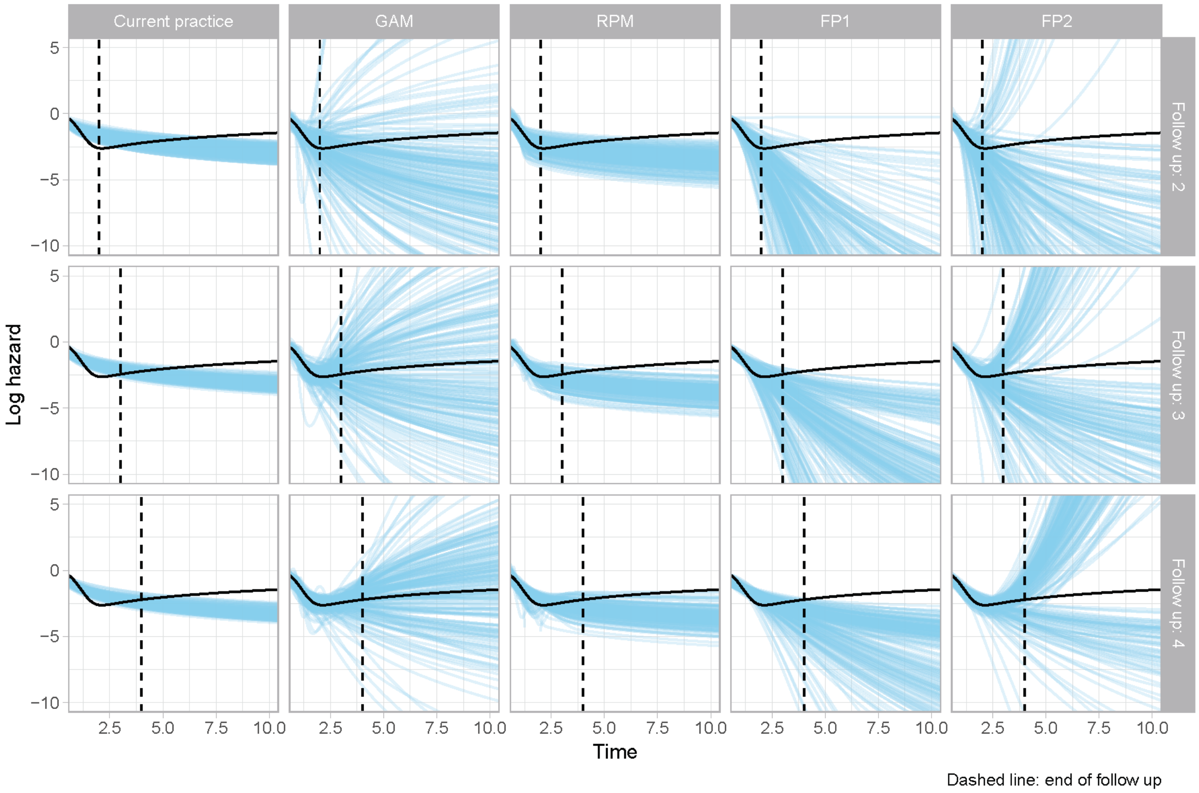
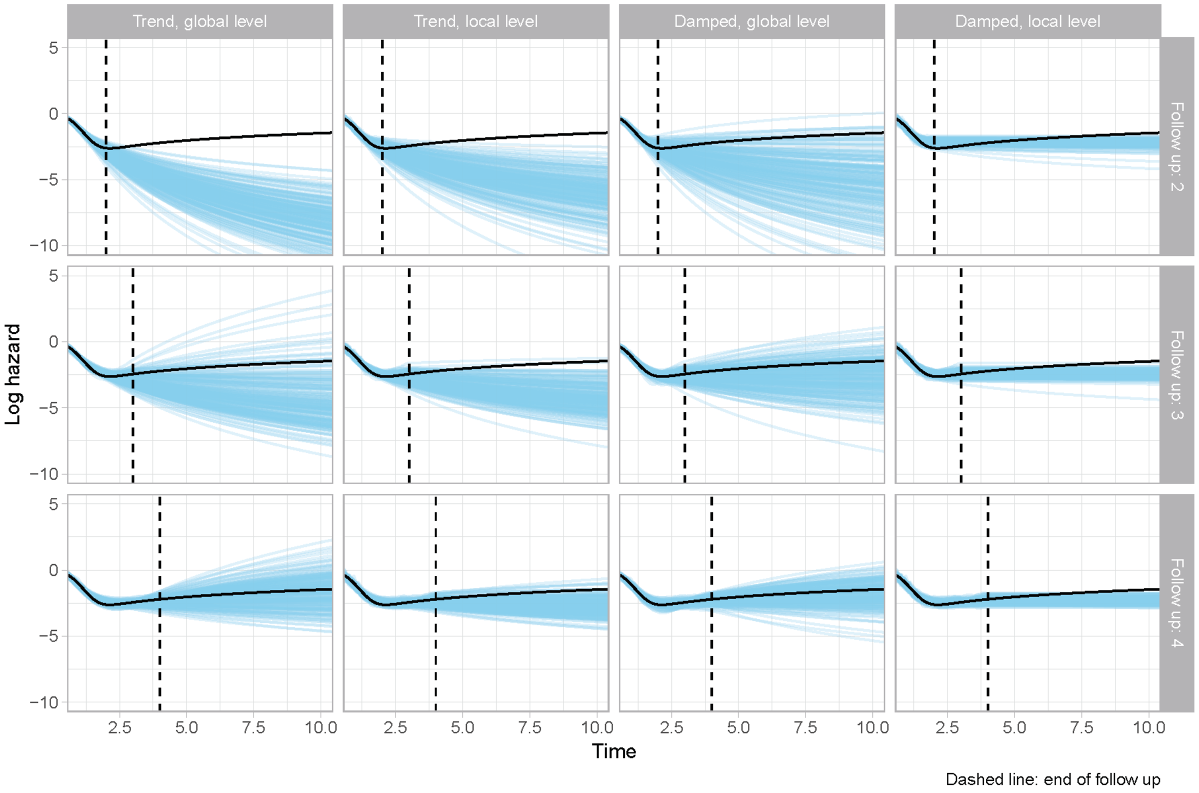


Figure 2.3: Model estimates of the log-hazard (blue lines) and true values (black lines), sample size = 600.

(a) Dynamic survival models.



(b) Current and emerging practice (excluding Gompertz).

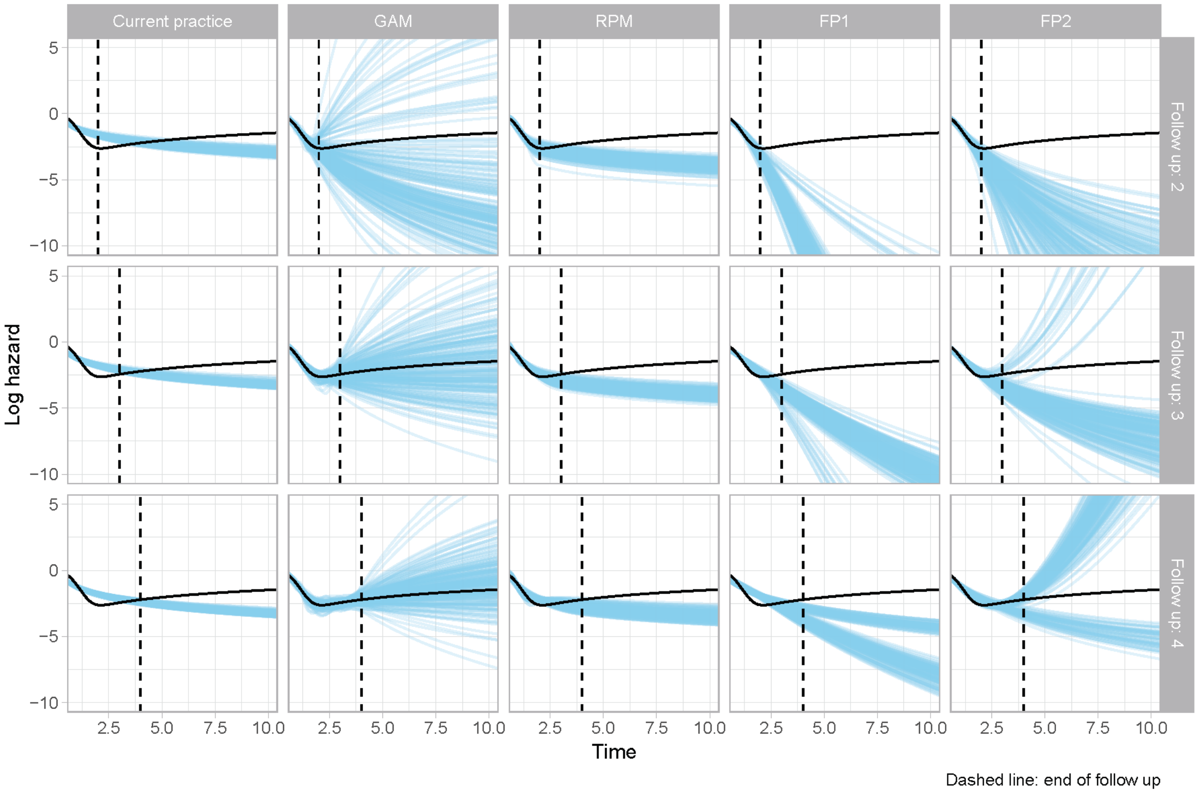
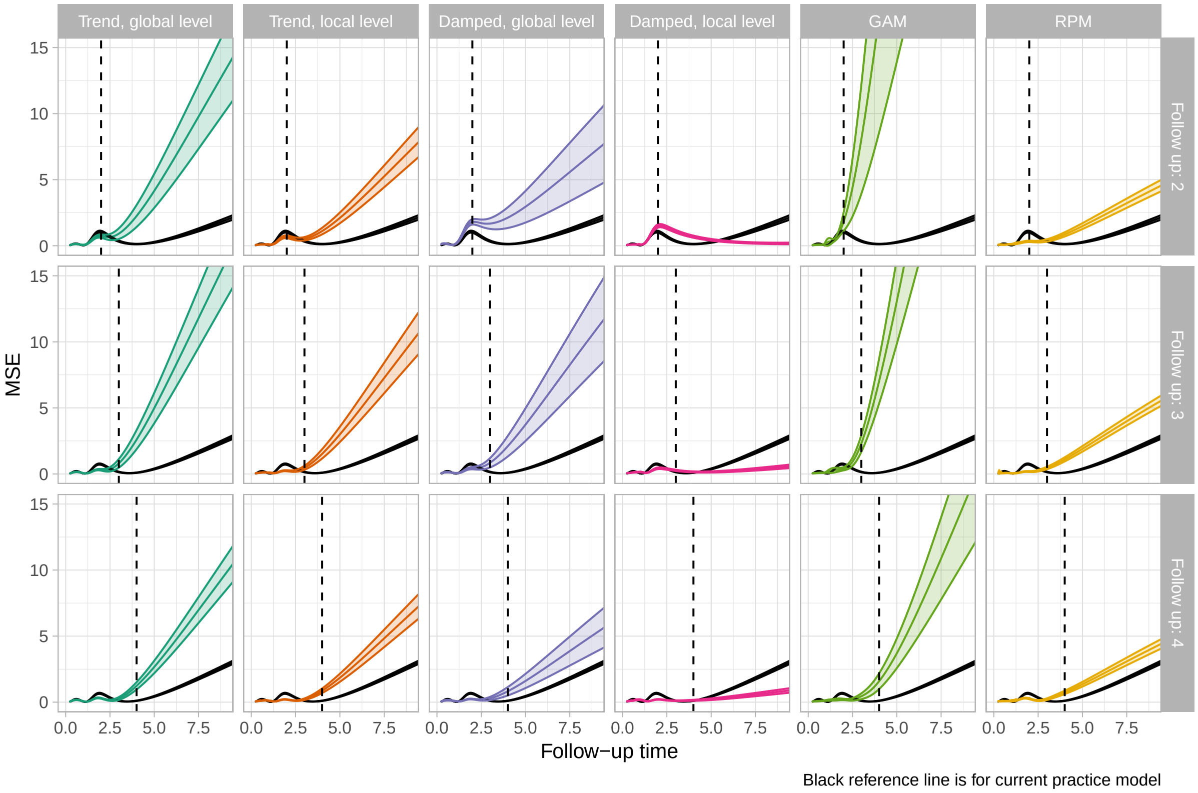


Figure 2.4: Mean squared error and bias values over time (within-sample and extrapolations), sample size = 100.

(a) Mean squared error (MSE).



(b) Bias.

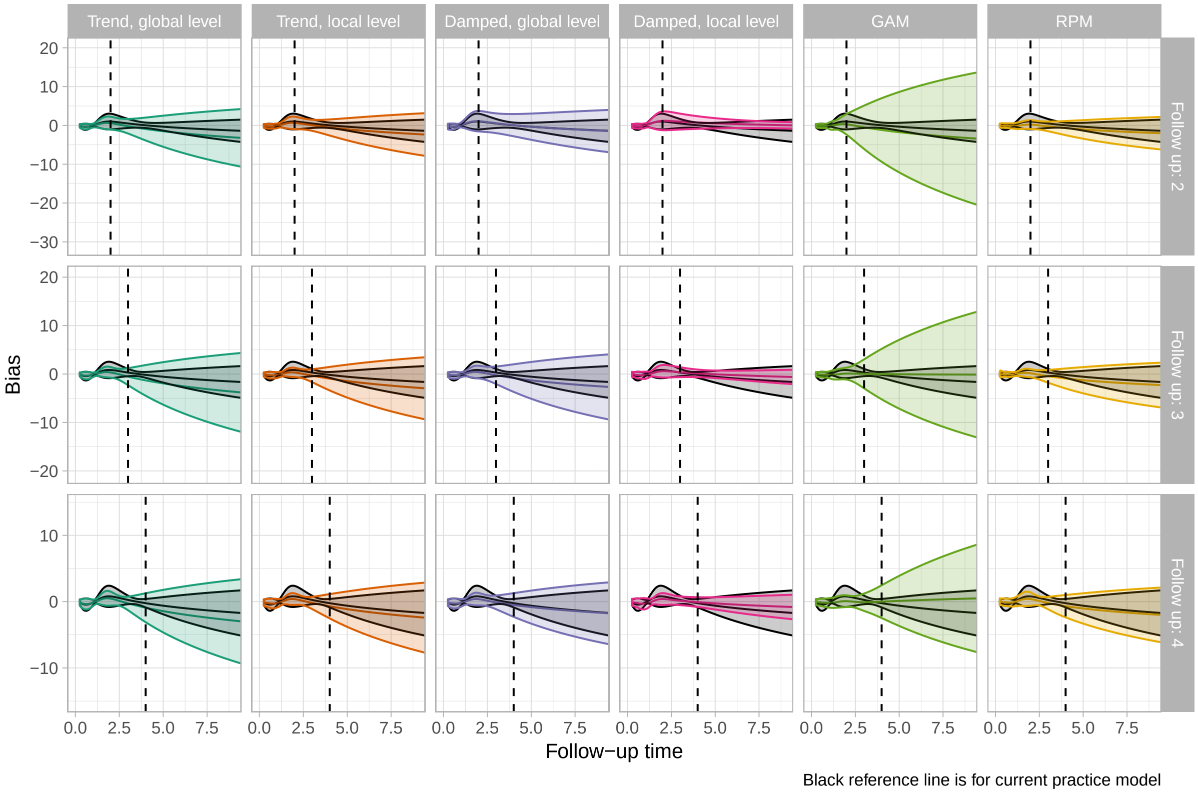
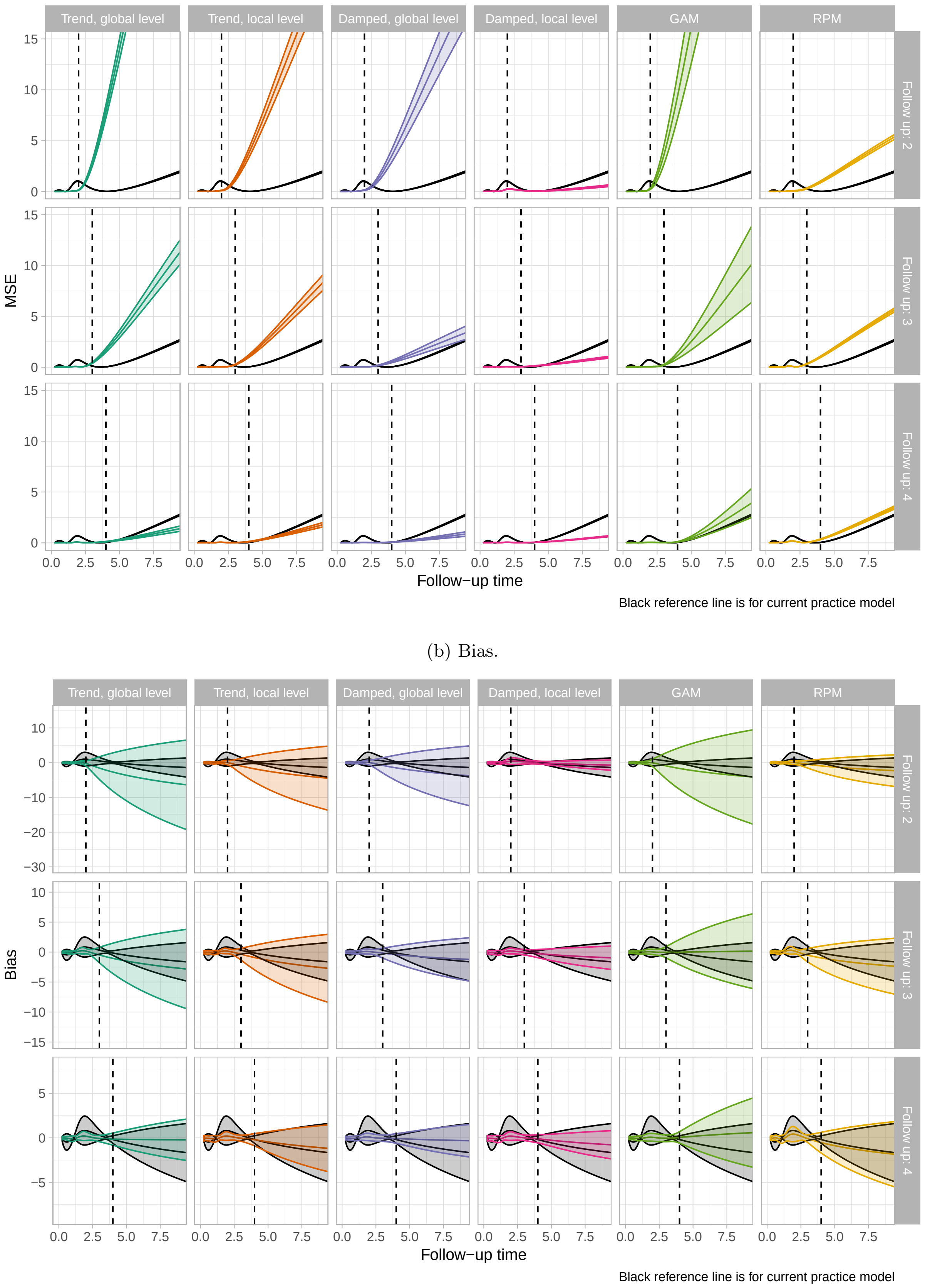


Figure 2.5: Mean squared error and bias values over time (within-sample and extrapolations), sample size = 600.



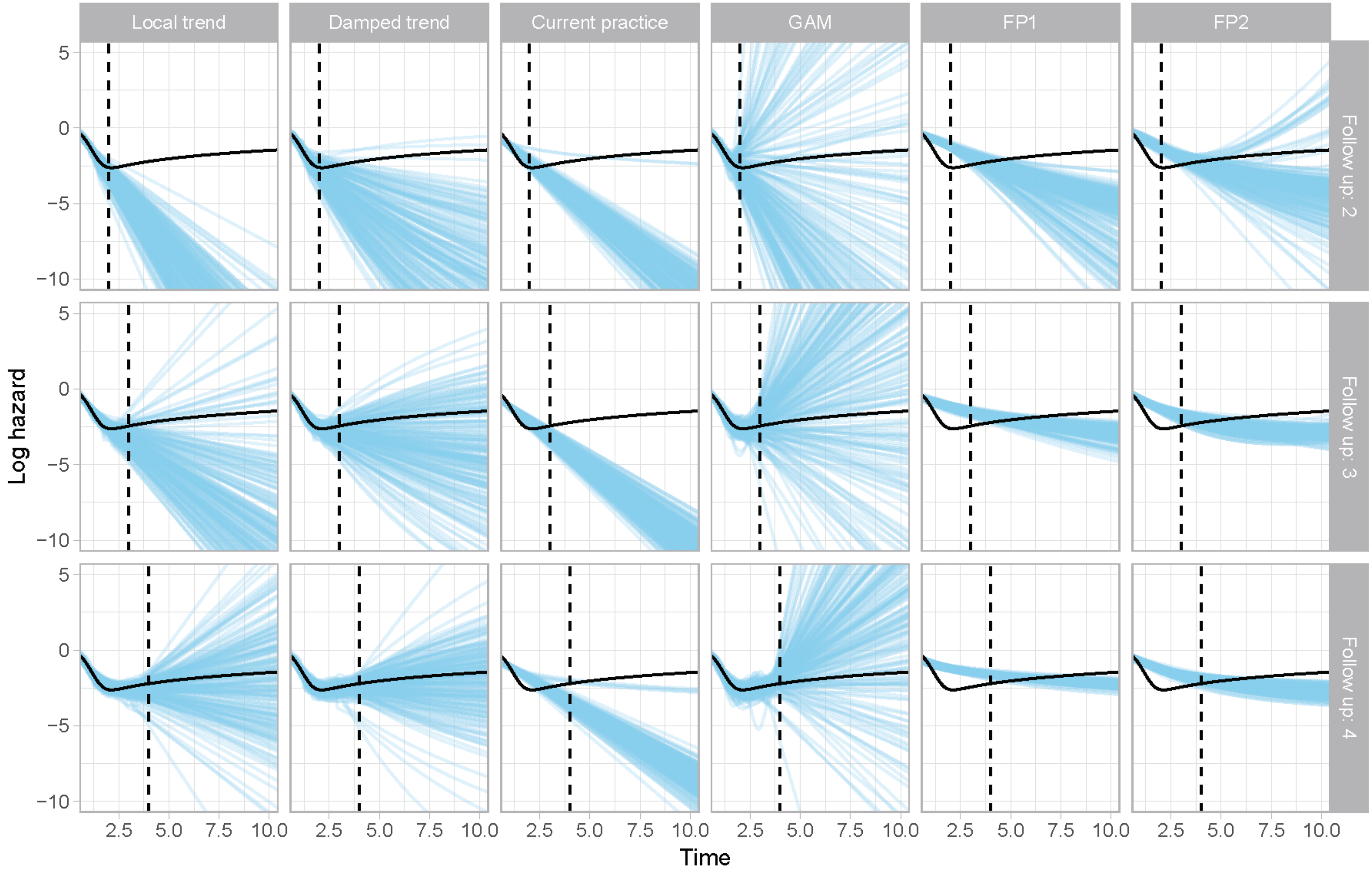
(a) Mean squared error (MSE).

Black reference line is for current practice model

(b) Bias.

Black reference line is for current practice model

Figure 2.6: Model estimates of the log-hazard (blue lines) and true values (black lines) for models using time as a covariate, sample size = 300.



Dashed line: end of follow up

2.5 Cure fraction simulation study

The following graphs are displayed:

- Visual goodness of fit (comparing model estimates with the truth) are provided in Figures 2.7 and 2.8 for sample sizes of 100 and 600, respectively. This uses the same specification for the models as for the main analysis.
- Estimates of mean-squared error and bias over time are provied in Figures 2.9 and 2.10 for sample sizes of 100 and 600, respectively. This uses the same specification for the models as for the main analysis.

Figure 2.7: Estimates of the log-hazard compared to the truth: cure models, sample size = 100

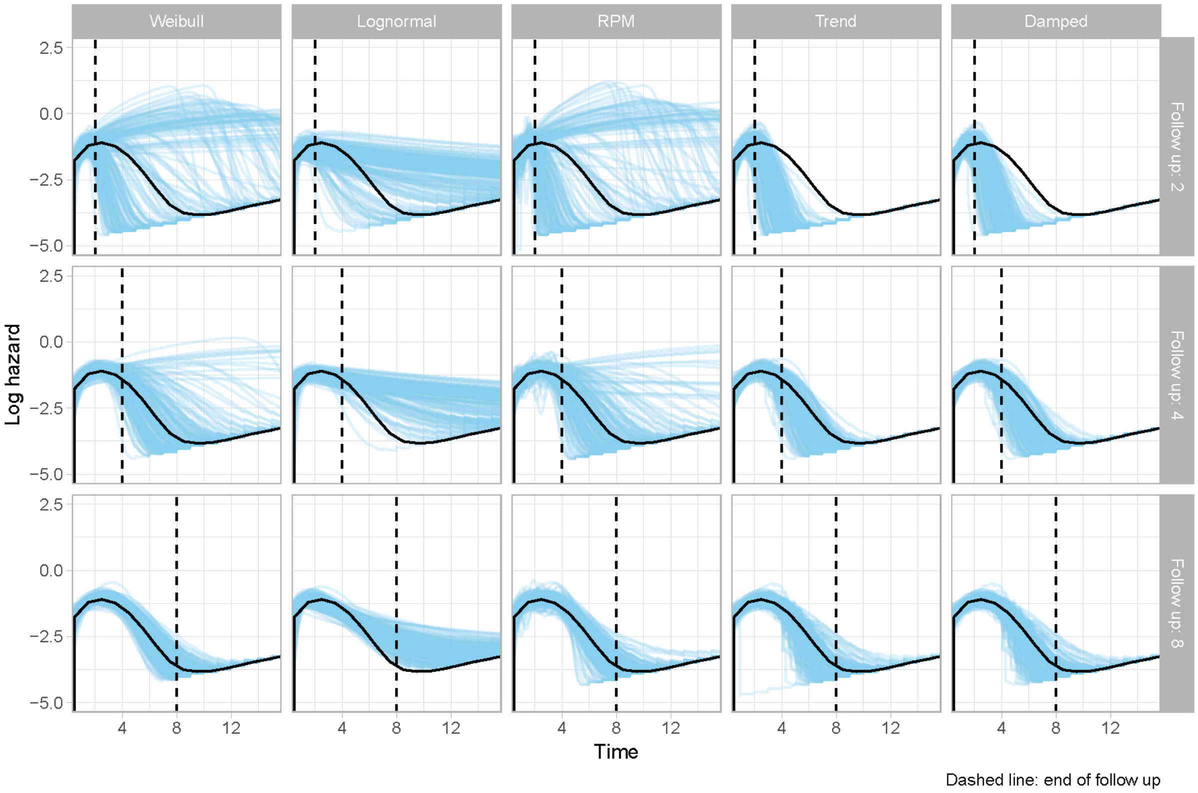


Figure 2.8: Estimates of the log-hazard compared to the truth: cure models, sample size = 100

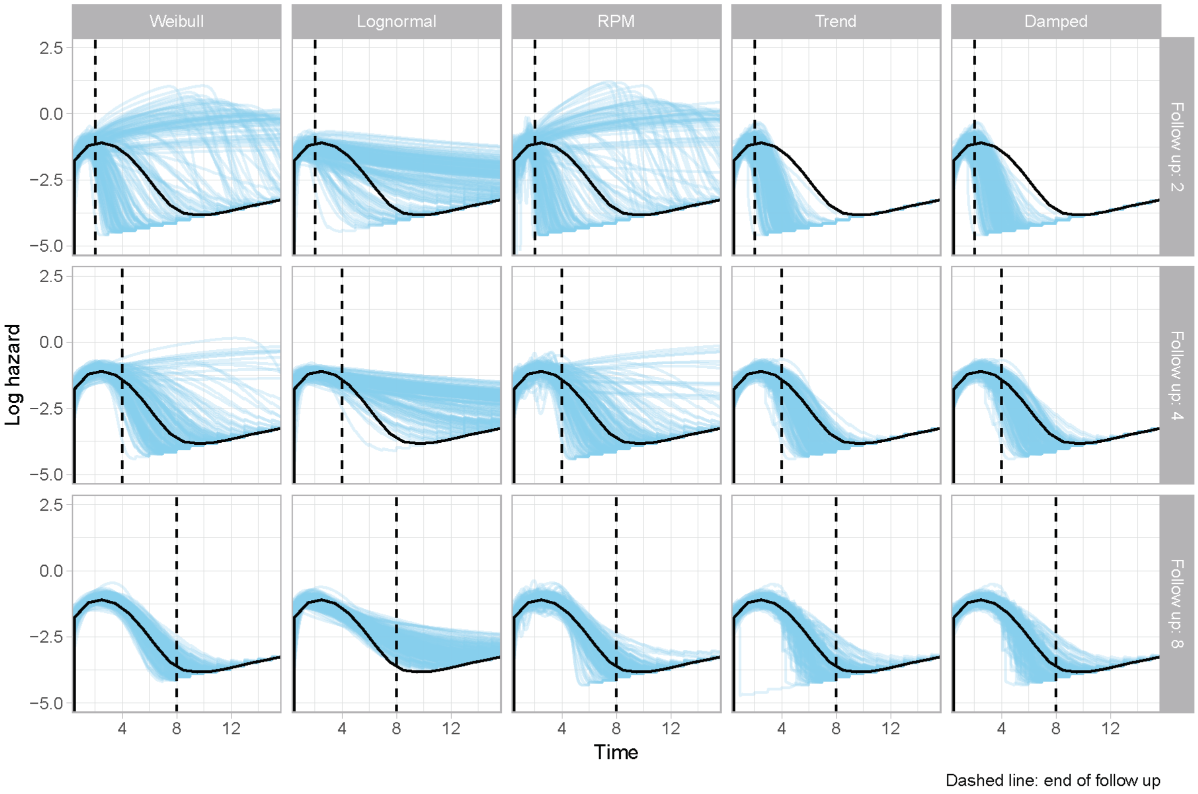
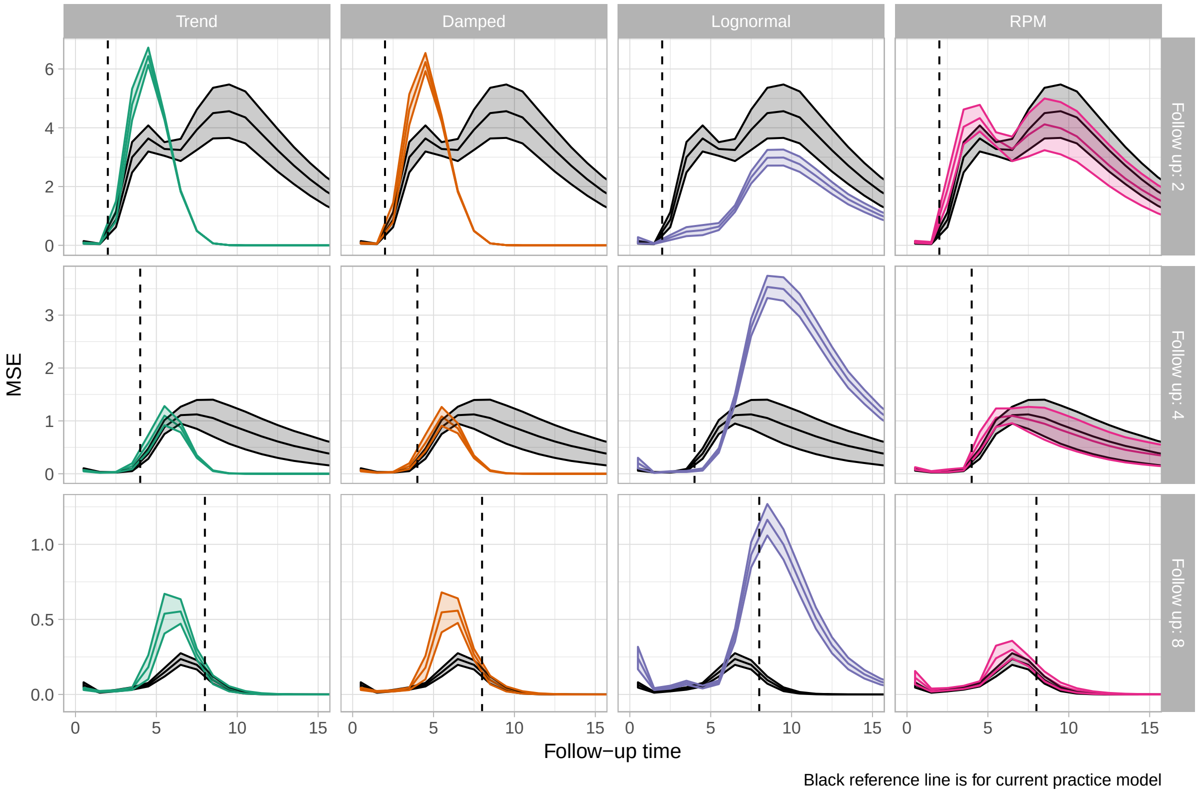


Figure 2.9: Mean squared error and bias values over time (within-sample and extrapolations), sample size = 100.

(a) Mean squared error (MSE).



(b) Bias.

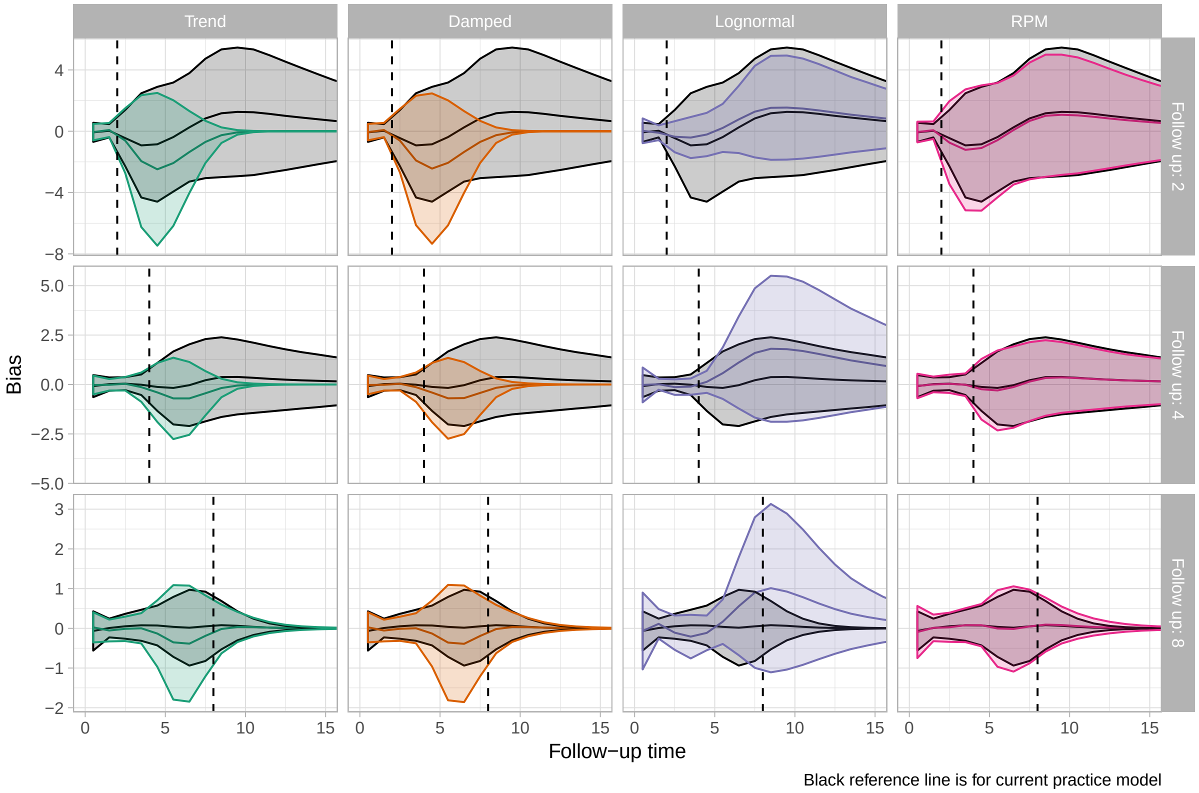
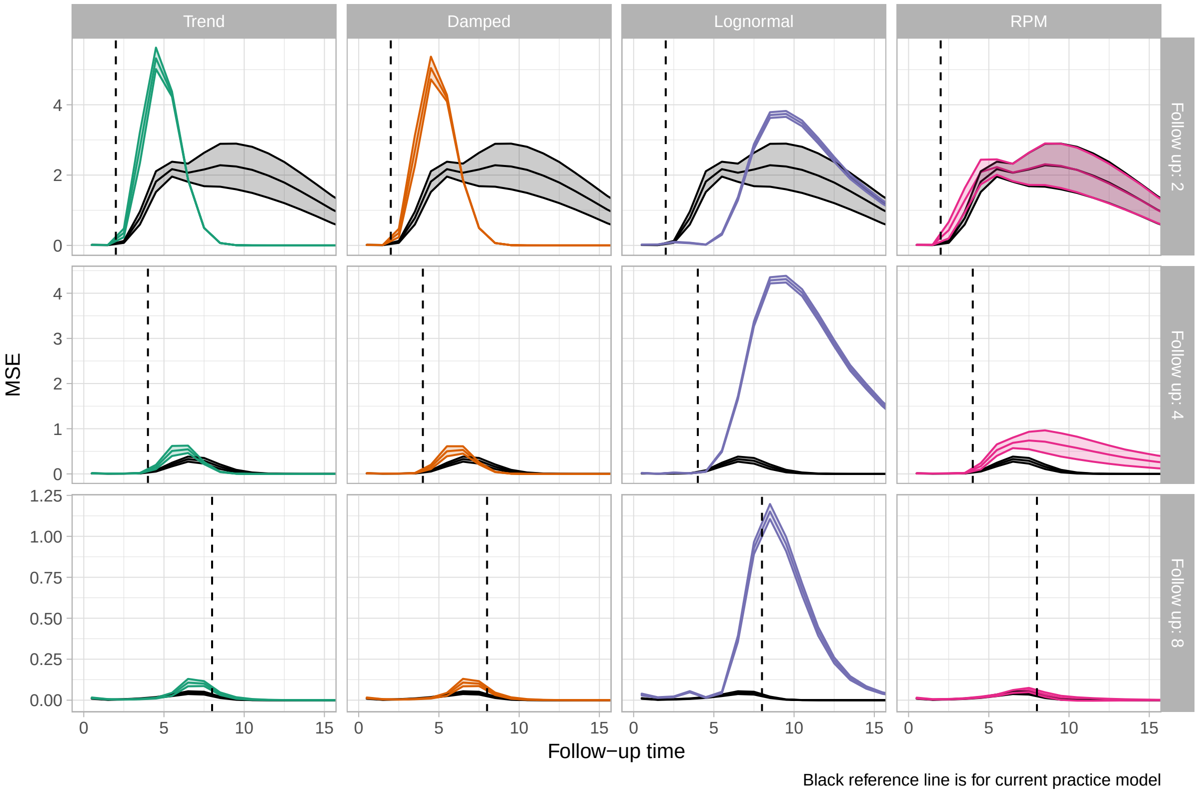
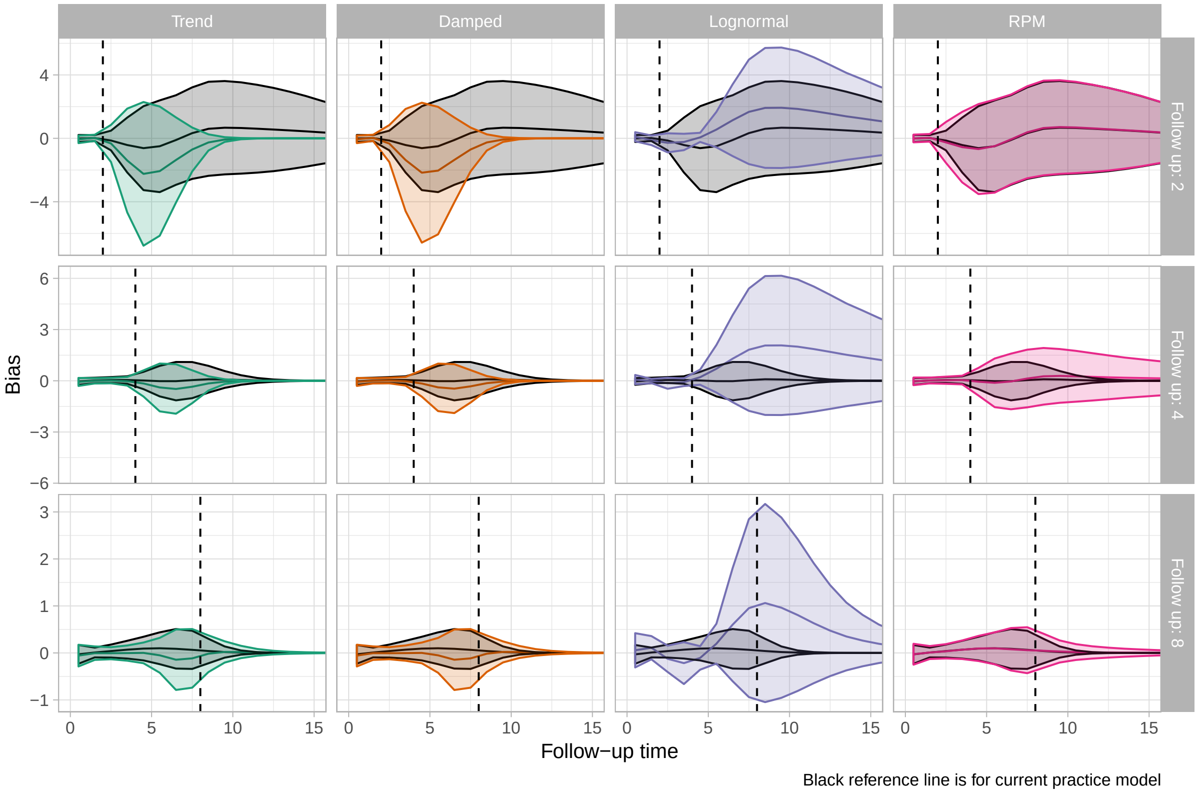


Figure 2.10: Mean squared error and bias values over time (within-sample and extrapolations), sample size = 600.

(a) Mean squared error (MSE).



(b) Bias.



Appendix 3

Generalised linear models for flexible parametric modelling of the hazard function

In many medical studies the outcome of interest is the time until an event occurs. Examples include mortality, disease progression, or hospital admission. To aid with decision-making the hazard function is estimated from parametric models. A prominent example is health technology assessment (HTA), which aims to quantify both the benefits to patients and the costs a healthcare system would occur if a treatment were funded [3]. To allow for fair comparisons across different treatments it is important that all relevant benefits and costs are quantified, which often requires use of a lifetime horizon [10]. However, time-to-event (TTE) data with complete follow-up are rarely available. As such, parametric models may be used to extrapolate model-outcomes to a lifetime, and hence obtain estimates of mean TTE (such as mean survival) [18, 241].

Standard one and two parameter models are available, including the exponential, Weibull, Gompertz, log-logistic and lognormal [17]. However, these models may not be sufficiently flexible to capture complex, time-varying hazards [98, 134]. In Section 3.1 we introduce generalised linear models (GLMs) and show that standard survival models may be expressed as GLMs. This provides insight into the limitations of the standard models: they all impose an assumption of linearity. More flexible parametric models that relax this assumption are required. A number of these have been proposed within the framework of GLMs and its extensions, but to-date they are seldom used to analyse TTE. These are described in Sections 3.2 and 3.3, with an overview in Section 3.4. An application of these is described in in Section 3.5, which demonstrates that the GLM-based models can provide superior within-sample estimates and more plausible extrapolations than standard survival models. Concluding remarks are provided in Section 3.6.

This manuscript has two aims. The first is to propose the use of GLMs for the analysis of TTE data. This includes flexible GLMs such as fractional polynomials (FPs) and restricted cubic splines (RCS), which are closely related to Royston-Parmar (R-P) models. The second aim is to present generalisations to GLMs: generalised linear mixed models (GLMMs) [246], generalised additive models (GAMs) [135] and dynamic generalised linear models (DGLMs) [174, 128].

3.1 Analysing time-to-event data within a generalised linear modelling framework

3.1.1 Standard survival models as linear models

The framework of GLMs extends (generalises) the standard linear model to response variables with distributions in the exponential family, including Normal, Poisson, Binomial, Gamma and Inverse Gaussian distributions [247]. An advantage of GLMs is that they provide a unified framework - both theoretical and conceptual - for the analysis of many problems, including linear, logistic and Poisson regression [146]. A random variable Y belongs to the exponential family of distributions if its probability density (or mass) function can be written as:

$$f(y_t;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{3.1}$$

where $a(y)$ and $d(y)$ are functions of the data, whilst $b(\theta)$ and $c(\theta)$ are functions of the distribution parameter θ and assumed to be twice differentiable. Equation (3.1) may also include other parameters, which are treated as nuisance parameters [146]. Examples for the Normal, Poisson and Binomial distributions are provided in Table 3.1. For these, $a(y) = y$.

Table 3.1: Normal, Poisson and Binomial distributions as members of the exponential family

Distribution	$b(\theta)$	$c(\theta)$	$d(y)$
Normal	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$	$\frac{-y^2}{2\sigma^2}$
Poisson	$\log\theta$	$-\theta$	$-\log y!$
Binomial	$\log(\frac{\pi}{1-\pi})$	$n\log(1-\pi)$	$\log\binom{n}{y}$

μ and σ^2 are a mean and variance, π is a probability, n the number of trials and $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ is the binomial coefficient.

For a TTE GLM, the observed outcome is the number of deaths during an interval: y_t. This is linked to the at-risk population at time t (denoted by τ_t) using a distribution from the exponential family. Use of the Poisson distribution assumes that $y_t = \tau_t \times \lambda_t$ where λ_t is the hazard at time t. Alternatively, use of the Binomial distribution assumes that $y_t = \tau_1 \times p_t$ where p_t is the cumulative probability of death. Model specification is [247]:

$$\text{Observation model: } E[y_t] = \mu_t \times \tau_t, \quad y_t \sim \text{exponential family distribution} \tag{3.2a}$$

$$\text{Response function: } \mu_t = h(\boldsymbol{x}_t^T \boldsymbol{\beta}) \tag{3.2b}$$

where $E[\cdot]$ denotes the expected value, bold font denotes a vector, and:

$\boldsymbol{\beta}$ is a vector of parameter coefficients to be estimated from the data,

$\boldsymbol{x}_t$ is a covariate, assumed known (with transpose $\boldsymbol{x}_t^T$), and

$h()$ is a one-to-one response function which maps the linear predictor ($\boldsymbol{x}_t^T \boldsymbol{\beta}_t$) to μ_t. Its inverse is known as the link function, and is denoted as $g()$.

Model parameters may be obtained via maximum likelihood estimation. The general expression for the logarithm of the likelihood is:

$$\log \mathcal{L} = \sum_{t=1}^{N} \mathcal{L}_t = \sum_{t=1}^{N} y_t b(\theta_t) + \sum_{t=1}^{N} c(\theta_t) + \sum_{i=t}^{N} d(y_t)$$

Where N is the number of time-intervals. For the Poisson and Binomial models, this becomes:

$$\text{Poisson: } \log \mathcal{L} = \sum_{t=1}^{N} \left[y_t \log(\theta_t) - \theta_t - \log(y_t!) \right] \tag{3.3a}$$

$$\text{Binomial: } \log \mathcal{L} = \sum_{t=1}^{N} \left[y_t \log\left(\frac{\pi_t}{1-\pi_t}\right) + n_t \log(1-\pi_t) + \log\binom{n_t}{y_t} \right] \tag{3.3b}$$

In summary, a GLM may be specified by three components:

1. The distribution from the exponential family, as defined in equation (3.1),
2. the response (or link) function, and
3. the covariate vector.

For survival analyses, options for μ_t include the (cumulative) survival function, its complement the (cumulative) failure function, the hazard function, and the cumulative hazard function - see [17, 124] for more details. Depending on the specification, we can express standard survival models as a linear model: $\mu_t = \beta_0 + \beta_1 x_t$. Table 3.2 provides these specifications. The log-logistic and lognormal distributions have a cumulative function as their outcome. It would not be sensible to model such an outcome as a constant value which demonstrates why there is no single-parameter special case of these models. In contrast, the Weibull and Gompertz distributions model a non-cumulative outcome, so it is possible to model this as a single value, resulting in the exponential model.

Table 3.2: Specification of standard survival models as generalised linear models

μ_t	**Distribution**	**Response function**	**Covariate**	**Model**
Hazard	Poisson	Exponential	None	Exponential
Hazard	Poisson	Exponential	Time	Gompertz
Hazard	Poisson	Exponential	Log(time)	Weibull
Cumulative Failure	Binomial	Logistic	Log(time)	Log-logistic
Cumulative Failure	Binomial	Inverse probit	Log(time)	Lognormal

An important aspect of survival data is that there is typically censoring of observations. Censoring occurs because for standard models the outcome is the time of the event occurring, and for some individuals the event is not observed (so it is censored). Within the GLM formulation, time changes from being the outcome to a covariate, so there are no censored observations.

Information on censoring is included by calculating the ‘at-risk’ sample, and including this information in the model. For models with a binomial distribution there is an explicit parameter for the sample size. For models with a Poisson distribution, information on the sample size may be incorporated as an ‘offset’ term.

3.1.2 Limitations with linearity

The assumption of linearity may not always be realistic. For example, for overall survival the hazard of all-cause mortality will increase over time due to patient ageing. In contrast, frailty effects may result in disease-specific mortality decreasing over time (as those with an increased hazard will die sooner, leaving those with a lower hazard). The impact of treatment on survival may also vary over time: there may be an initial elevated risk of death due to adverse events; treatment-related toxicities may increase other-cause mortality over time, treatment stopping rules and trial inclusion criteria may have an effect [19]. These considerations motivate the need for more flexible survival models, which are considered within the GLM framework in the next two Sections.

3.2 Relaxing the assumption of linearity

We briefly describe flexible models that may be applied to survival data within a GLM framework, more details are provided in the Appendix. Without loss of generality, y is used to denote either a random variable or the observed data.

3.2.1 Fractional polynomials

FPs represent the outcome as a sum of polynomial terms; increasing the number of terms (the order of the FP) increases the flexibility of the model. A closed-test procedure may be used to identify the order. For a single variable, an i^{th} order FP is defined as:

$$E(y_t) = \text{FP}(i) = \beta_0 + \sum_{j=1}^{i} \beta_j x^{p_j} \tag{3.4}$$

where the set of powers p_j is pre-specified, and may include fractional powers (hence the name fractional polynomials). FPs include linear models as special cases, so depending on specification may include one of the standard models from Table 3.2. Some limitations with FPs are that they may not have sufficient power to detect non-linearity, and they can be sensitive to extreme values in the data. This sensitivity occurs because FPs are *global* models: $\boldsymbol{\beta}$ values are assumed to be constant over time.

3.2.2 Restricted cubic splines and Royston-Parmar models

A cubic spline represents a continuous function as a series of piecewise cubic polynomials [124], hence relaxing the assumption of global time effects. Model flexibility is based on the number of piecewise intervals (equivalently, the number of 'knots'). For extrapolation, the cubic polynomial from the last interval may be used, or it may be restricted to a linear function: this latter assumption results in an RCS. An example specification is provided in the Appendix.

R-P models use RCSs, but not in the GLM framework. Typically the outcome is the log cumulative hazard, which is monotonic. However, model estimates are not guaranteed to be monotonic, so implausible values may result.

As they are not global models, splines may over-fit local 'noise' in the data [205], and there is in general no closed test procedure for choosing between different models.

3.3 Extensions to the generalised linear model

This section provides a brief overview of extensions to GLMs, with more details in the Appendix.

3.3.1 Generalised linear mixed models

A GLMM extends the GLM by incorporating random effect terms, which can help to quantify the impact of unmeasured covariates and provide more realistic estimates of uncertainty. An example of an FP(2) with a random-effect (denoted by b_t) is:

$$E(y_t) = \text{FP}(2) = \beta_0 + b_t + \beta_1 x^{p_1} + \beta_2 x^{p_2}, \quad b_t \sim N(0, \psi^2)$$

GLMMs are also referred to as *frailty* models [107]. In theory, any GLM may be extended by adding a random term as shown above. The main limitation with GLMMs is that as the random effects are not observed, there may be difficulties in model specification and parameter estimation.

3.3.2 Generalised additive models

A GAM is a GLM in which one or more of the covariates are modelled as a set of *basis* functions [248]. For example, a univariate GAM is defined as:

$$E(y_t) = \sum_{j=i}^{q} b_j(t)\beta_j = f(t)$$

Where $b_j(t)$ is the jth basis function, and q is the dimension of the basis function. Higher values of q result in more flexible models. Both FPs and RCSs may be viewed as GAMs. The main extension provided by a GAM is that model complexity is penalised during parameter estimation (via shrinkage of the $\boldsymbol{\beta}$). GAMs with a cubic spline basis have theoretical justification as being approximate 'smoothest interpolators' [135] - see the Appendix for more details. Limitations of GAMs will depend on the basis function used. For example, if a spline is used, the limitations of these will still apply.

3.3.3 Dynamic generalised linear models and dynamic survival models

In a DGLM model coefficients ($\boldsymbol{\beta}$) are allowed to vary over time. When applied to TTE data, DGLMs are known as *dynamic survival models* (DSMs) [144]. Specification is (compare with Equation 3.2a):

$$\text{Observation model: } E[y_t] = \mu_t \times \tau_t \quad y_t \sim \text{exponential family distribution} \tag{3.5a}$$

$$\text{Response function: } \mu_t = h(\boldsymbol{x}_t^T \boldsymbol{\beta}_t) \tag{3.5b}$$

$$\text{Transition model: } \boldsymbol{\beta}_t = F\boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t \tag{3.5c}$$

$$\text{Initial conditions: } \boldsymbol{\beta}_0 \sim MVN(\boldsymbol{b}_0, \boldsymbol{Z}_0) \tag{3.5d}$$

where MVN denotes a multivariate Normal distribution, F is a function describing how the coefficients evolve over time, and ζ_t is an error term - see the Appendix for further details. DGLMs may be viewed as combining GLMs with time-series methods. In particular, parameter estimates may be based on minimising the error of within-sample extrapolations. This makes these models particularly appealing when the primary objective of the analysis is extrapolation. The main limitations with DGLMs are identifying suitable initial values, and convergence of algorithms to estimate model coefficients [143, 144].

3.4 Theoretical comparison of approaches

Five different modelling approaches were considered: FPs, splines, GAMs, GLMMs, and DGLMs. The frailty terms from a GLMM may be combined with either of the other four models. The following prompts are provided to aid with choosing between the different approaches.

What is the primary objective of the analysis? If the main objective is in generating extrapolations, this implies the use of a DGLM, as this is the only one of the models for which parameter estimation is based on minimising forecasting error. If instead the main objective is to provide estimates of the observed data, then any of the approaches may be used.

Fractional polynomials or spline-based models? Spline-based models may be preferred on theoretical grounds, as being approximate smoothest interpolators, whilst there are a number of limitations with the use of FPs (see the Appendix). This suggests the use of a spline-based model in preference to an FP within a GLM framework, with the latter as a form of sensitivity analysis.

To penalize during or after estimation? Parameter estimation with a GAM automatically penalises for model complexity, which helps to avoid over-fitting. Alternatively, information criteria may be used. There are a number of different information criteria that could be used, whereas GAMs have a specific objective function. The choice between these is likely to be study specific: sometimes there may be good reasons to use a specific information criteria, whilst in others the more automated approach of a GAM may be preferred.

For both approaches it is not possible to use significance tests to choose between model specifications.

Are there any subject matter considerations? For example, there may be reason to believe that there are important unmeasured confounders, which suggests incorporating random effects. Or it may be thought that there will be important local fluctuations in this hazard, which suggests the use of either a spline or dynamic model in preference to the global FPs.

3.5 Empirical comparison of approaches

3.5.1 Dataset

We use a freely available dataset to demonstrate both the limitations of assuming linearity and the use of more flexible models. Analyses were performed in R; the code used is available as supplemental material. Hence the case-study is fully reproducible.

The data are on the survival of individuals following a diagnosis of breast cancer, and from a study conducted by the German Breast Cancer Study Group [136, 100]. Individuals with primary node positive breast cancer were recruited between July 1984 and December 1989. Events are defined as either cancer recurrence or death (from any cause). Data are available for 686 individuals, of which 299 experienced an event during follow-up. The maximum follow-up was 7.28 years, with mean follow-up of 3.08 years. Use of GLMs requires that individual-level data are restructured in the form of life tables. Samples of the individual-level data and the corresponding (monthly) life table are provided in Tables 3.3 and 3.4, respectively. For Table 3.3, an event indicator of one denotes that an event occurred (otherwise the indicator is zero, and the outcome is time to censoring).

As described in Section 3.1.2, the assumptions of linearity imposed by standard two-parameter survival models may be unrealistic. To highlight this, we show model estimates against the observed data in Figure 3.1 for each model (the one-parameter exponential model is not shown as it would only be appropriate if both the Weibull and Gompertz estimates had no slope). The

Table 3.3: A sample of the breast cancer data

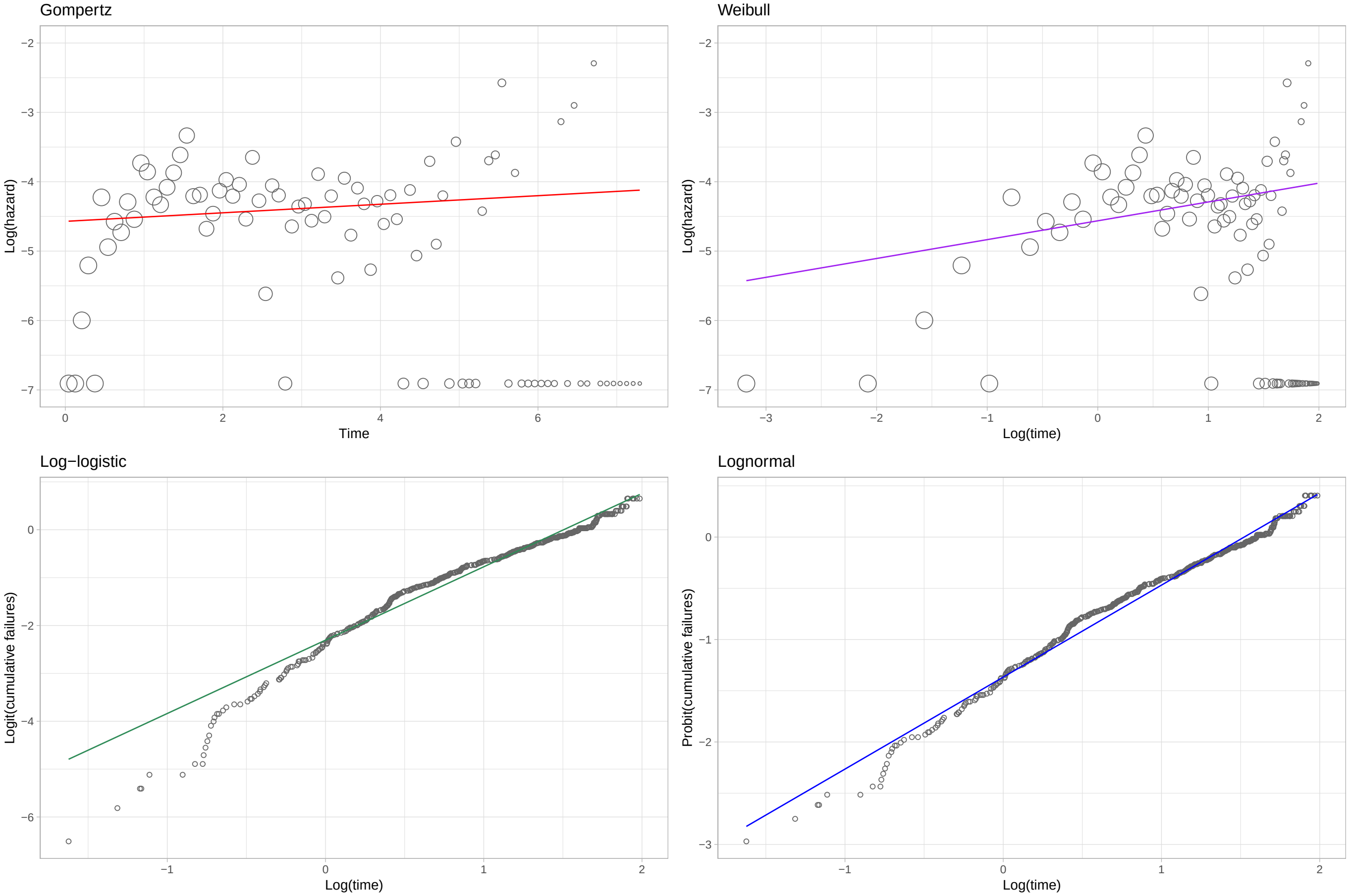
Patient ID	Outcome time (years)	Event indicator
1	0.0219	0
⋮	⋮	⋮
15	0.1973	1
⋮	⋮	⋮
220	1.9562	1
221	1.9644	0
⋮	⋮	⋮
678	6.7288	1
⋮	⋮	⋮
686	7.2849	0

Table 3.4: Data from Table 3.3 restructured for Poisson regression

Month	Sample size	Events (μ)	Censorings	At risk (τ)	Hazard (λ)
(0, 1)	686	0	7	682.5	0
(1, 2)	679	0	3	677.5	0
(2, 3)	676	1	4	674	0.001
⋮	⋮	⋮	⋮	⋮	⋮
(22, 23)	477	5	3	475.5	0.011
(23, 24)	469	7	4	467	0.015
(24, 25)	458	8	12	452	0.018
⋮	⋮	⋮	⋮	⋮	⋮
(87, 88)	1	0	1	0.5	0

specification of the x and y axis is such that the model estimates form a straight-line. Figure 3.1 shows that the linear estimates generally provide a poor visual description of the data, with the best description arising from the lognormal model.

Figure 3.1: Breast cancer case-study: observed and modelled hazard



3.5.2 Methods

We considered five broad classes of model:

FP models. We considered FP(2) models, with the complexity of the chosen model based on the closed-test procedure, and the chosen powers based on minimising AIC.

Generalised linear mixed models. We fit FP models as described above, but we also included frailty terms.

Spline-based models. Both RCS models and GAMs were considered. For the RCS model between 1 and 5 internal knots were considered, with the choice based on minimising AIC. For the GAM we considered two approaches to selecting the dimension of the basis function: one used a fixed (arbitrary) value of 11 (v1), the other was based on minimising AIC (v2). These two approaches were considered as some penalisation for over-fitting is included during model-fitting, so it is unclear if model choice based on AIC is required. For all models, the knots were placed at equally-spaced percentiles of the observed uncensored death times [136].

Dynamic models. We examined three specifications: local-level, local-trend, and local-level with global trend. There was no need to base model choice on minimising AIC (as the data used to estimate the model parameters are separate to the objective function, which is based on minimising one-step ahead forecasts).

Standard survival models. Eight survival models were considered: exponential, Weibull, Gompertz, gamma, log-logistic, lognormal, generalised gamma, and generalised F. Results are displayed for the three best fitting models (based on AIC). Note that the generalised gamma and generalised F models have three and four parameters respectively, so are more flexible than the standard survival models of Table 3.2.

The above choice of models was designed to be representative of the variety of different approaches possible, but not exhaustive. All of the models used the natural logarithm of time as the only covariate of interest (with the exception of the Gompertz, which uses time). All of the GLM-models assumed a Poisson distribution with an exponential response function.

3.5.3 Goodness of fit

Goodness of fit (GoF) measures how well the statistical model describes the observed data. It should be distinguished from predictive ability, which measures how well the model predicts external data (such as future observations). One measure of GoF is Akaike's information criterion (AIC), which is defined as:

$$-2\log\mathcal{L} + 2k \tag{3.6}$$

where $\mathcal{L}$ is the model likelihood and k is the number of parameters in the model [201]. Because the likelihood is multiplied by a negative number, lower AIC values are to be preferred.

A number of variants on AIC have been proposed [201, 42]. An empirical study by Hyndman and colleagues [42] compared five GoF measures, and noted that they all performed similarly. Further, Burnham and Anderson note that the AIC has strong theoretical motivation [201], whilst Jackson and colleagues note that the AIC is preferable when models are used to represent complex phenomena (such as survival processes) [249]. Due to having both empirical and theoretical support, the AIC shall be used in this manuscript. Any GoF measure should be used in combination with subject-matter considerations. In addition, estimates of the hazard function were visually compared to the observed hazard function.

The AIC measures GoF to the observed data. It is unknown if models with a good within-sample fit provide good extrapolations [124]. To measure the extrapolation performance of the models we split the dataset into two parts. The first part considered events occurring within the first three years, censoring all events after three years (half of the sample were at-risk of an event at three years). Extrapolation performance was defined as the sum of squared errors (SSE) between the model-estimate of the hazard and the observed hazard (calculated for monthly intervals) for the remaining follow-up:

$$\left(\hat{\lambda}_t - \lambda_t\right)^2, t \in \{37 \text{ to } 88 \text{ months}\} \tag{3.7}$$

3.5.4 Results

Table 3.5 provides GoF values for each model and estimates of lifetime mean life expectancy. Two AIC values are provided: one using the entire dataset, the other using the first three years.

The number of parameters is provided as a measure of model complexity: the two GAMs do not have an integer number of parameters, as parameter effects are shrunk during model estimation. Plots of the estimated hazard function for each model are displayed in Figure 3.2 for the observed data. Corresponding extrapolations are given in Figure 3.3. As the best-fitting two-parameter standard survival model (based on all the available data), the lognormal is provided as a black reference line on all panes.

Within-sample goodness of fit

All of the more flexible models provide lower AIC values than the lognormal, although in general differences between values are small, and cannot be tested for statistical significance. Visually, all of the models provide a good fit to the observed data in Figure 3.2, although there is variation in the degree to which local fluctuations are captured.

Of the 11 models, the lowest AIC values arose from two DSMs. However, the third DSM had the highest AIC of all the flexible models. This suggests that the extension to dynamic models can lead to an improved GoF, but there is no guarantee that this will always occur. The next best AIC values arose from the three spline-based models, which all had very similar GoF. However, the two approaches to GAM estimation did result in markedly different models: the one with automated fitting was more complex (with almost three times as many parameters) than the one based on minimising AIC, whilst also providing a better absolute fit (based on log-likelihood).

Of the three standard survival models, the two generalised models (gamma and F) both provided similar GoF, and both improved on the two-parameter models. Fit for the two FPs was similar to that for the generalised gamma and generalised F survival models, and lower than that for the spline-models. The inclusion of random effects had a negligible impact on the AIC.

Flexible parametric modelling of the hazard provides insight into how it varies over time. The GAM (v1) and DSMs were slightly better at capturing local fluctuations in the hazard rate. This is most notable at approximately 1 and 1.5 years. However, as the most flexible models considered, there is a danger that these local fluctuations represent noise. If this is the case then the best-fitting models may be over-fitting the data, with no guarantee that this will lead to improved extrapolations.

Extrapolation goodness of fit

When fitting the 11 models to the first three years, the ranking of the models was generally the same as for the full dataset, with the local level model providing the lowest AIC, and the lognormal one of the highest. An exception is the DSM with drift, which changes from having the second lowest AIC to the second highest. GoF to the observed data did not predict extrapolation performance. For example, the lognormal and local trend models both had the highest AIC values but the lowest SSEs. As with the AIC values, in general there was little difference between SSE values. An exception is the DSM with a drift, which provided poor extrapolations as it predicted an increasing trend.

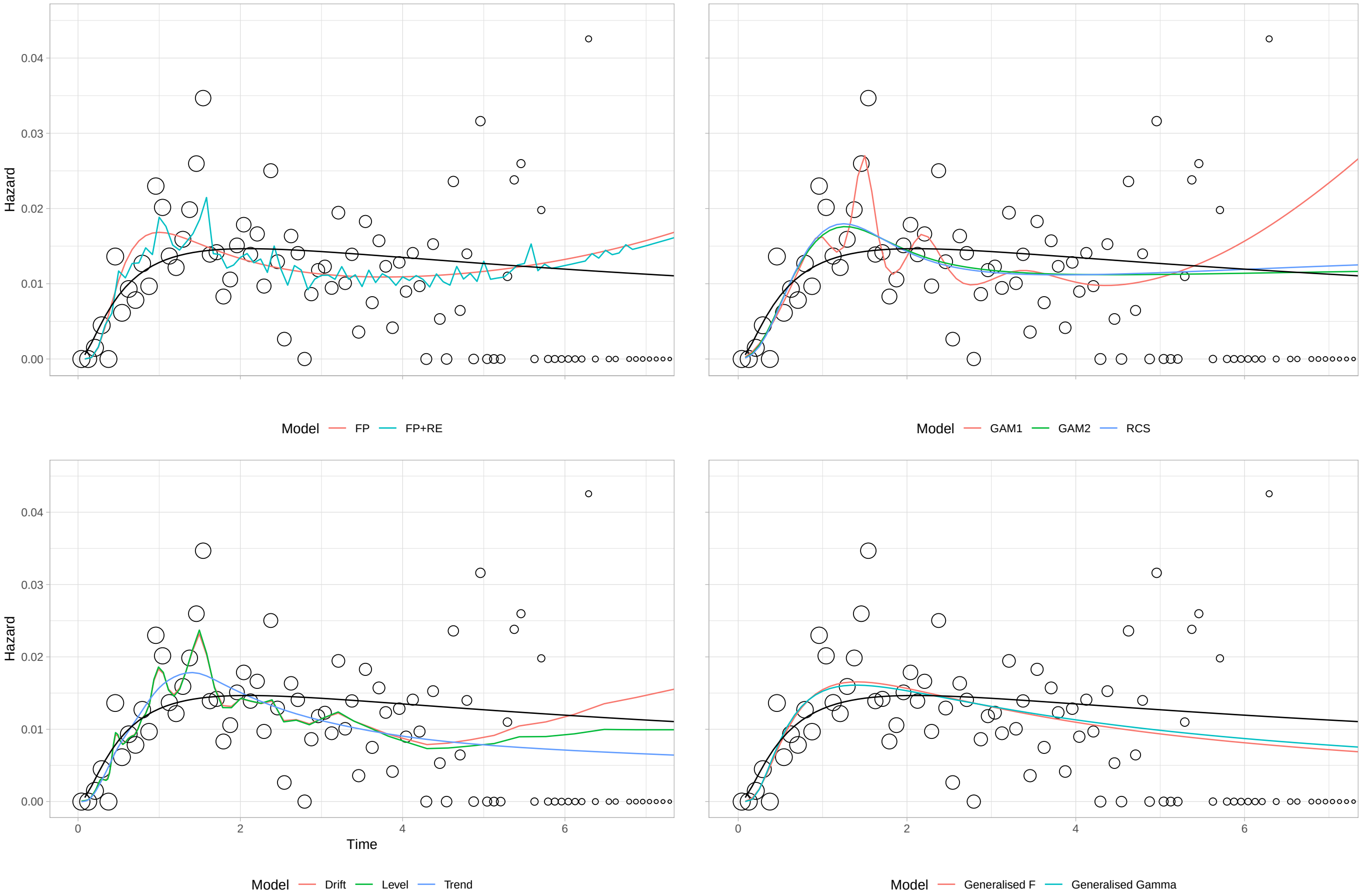
In general the results in Table 3.5 demonstrate that there is little difference between the competing models, both for within-sample and extrapolated GoF. However, Figure 3.3 shows that resulting extrapolations (beyond the full data follow-up) can vary markedly by model. Differences begin at about four years, and are likely to be due to the small patient numbers. For example, at five years the sample size at risk is 113, at six years it is 34 and at seven years it is three. When choosing between the models, it is very important to assess the plausibility of the extrapolations with clinical experts, noting the outcome definition used. For this case-study the mean age of the sample is 53 years and the outcome is either cancer recurrence or death from any cause. The mean survival for German women of this age was 32.6 years in 2000 (the oldest year for which there is data). This acts as an upper-bound on the likely survival of this sample, as women with breast cancer are likely to have worse survival than the age-matched general population, and cancer recurrences would further reduce the estimated survival. Of the 11 models considered, only the four which predicted an increasing extrapolated hazard (DSM with drift, GAM with default settings, both FPs) gave a lifetime mean survival less than this.

Table 3.5: Breast cancer case-study: log-likelihood and information criteria for the models

Model	Log-likelihood	Parameters	AIC: full Data	AIC: years 1-3	SSE: years 4-7	Life Expectancy
Local level	-142.72	3	291.45	168.48	3.84	37.62
Local level with drift	-142.09	4	292.19	180.25	18.58	23.41
GAM v2	-150.63	3.84	308.94	172.08	4.01	37.12
RCS	-150.55	4	309.10	172.12	4.05	35.46
GAM v1	-144.05	10.66	309.42	173.89	3.81	14.13
Generalised Gamma	-153.03	3	312.06	175.31	3.78	43.40
FP with random effects	-152.13	4	312.27	173.54	4.25	15.70
FP	-153.42	3	312.84	172.51	4.29	15.40
Generalised F	-152.97	4	313.94	174.40	4.01	43.87
Local level local trend	-152.36	5	314.71	180.68	3.76	41.61
Lognormal	-157.55	2	319.11	179.42	3.73	40.64

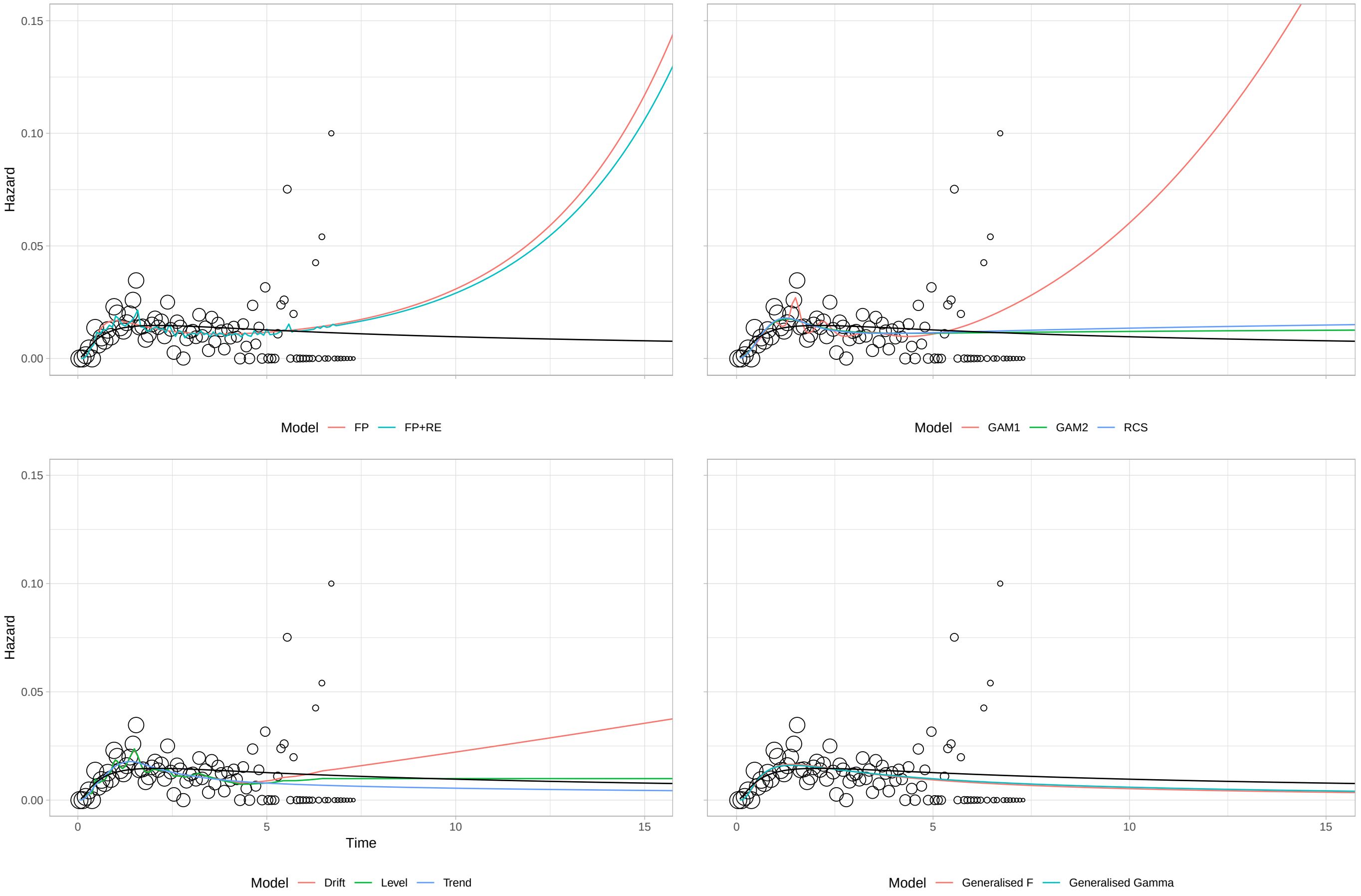
AIC: Akaike's information critera. FP(2): Second-order fractional polynomial. SSE: Sum of squared errors (× 10,000)
For derivation of SSE values see section 3.5.3

Figure 3.2: Breast cancer case-study: observed and modelled hazard



FP: fractional polynomial. RE: random effects. RCS: restricted cubic spline. GAM: generalised additive model. Gen: generalised. Hollow circles represent observed data; sizes are proportional to the denominator. For all panes the lognormal is included in black. Three observations are removed: see Figure 3.3 for these.

Figure 3.3: Breast cancer case-study: extrapolated hazards



FP: fractional polynomial. RE: random effects. AR: autoregression. RCS: restricted cubic spline. GAM: generalised additive model. Gen: generalised. For all graphs the lognormal is included in black.

3.6 Discussion

A wide variety of flexible parametric models may be used to analyse and extrapolate TTE data within a GLM framework, along with its extensions to GAMs, GLMMs and DGLMs. These include FPs, spline based models and DSMs. An advantage of the GLM-based models over standard survival models is that they can be made arbitrarily flexible as required to match the complexity of the observed hazard function (for example, increasing the order of an FP or the number of knots in a RCS). In contrast, to obtain more complex standard survival models, different specifications are required (such as moving from a Weibull to a generalised gamma model). Further, two of the GLM-extensions (GAMs and DGLMs) penalise for over-fitting as part of parameter estimation [143, 135], thus removing much of the subjectivity over model choice. To our knowledge, this is the first time that all of these approaches have been compared at both a theoretical and an applied level, with recommendations to aid in choosing between the models.

The case-study demonstrated that it is straight-forward to perform a TTE analysis within a GLM framework and that results are at least as good as, and often superior to, those from standard survival models. However, differences in GoF were typically small, and in this example there was no relationship between within-sample GoF and extrapolation performance. A strength of the case-study is that we considered a variety of different statistical models, some of which are currently infrequently used in survival analyses [18, 144]. The fully reproducible nature of the case-studies shall help to increase the uptake of these more advanced methods.

There were marked differences in the extrapolations from each model, and hence estimates of lifetime mean survival. Using external evidence, only the extrapolations from one each of the DSMs and GAMs along with both FPs were plausible, whilst the best three standard survival models all provided implausible extrapolations. This highlights a further benefit of the GLM-approach, as it increases the potential to identify models which simultaneously provide good within-sample fit and plausible extrapolations. Formally incorporating such evidence is an important area of on-going research [79, 22]. However, this task is often non-trivial. For example, external datasets may exist but they may not be fully generalizable to the decision problem. This could be due to differences in the patient population, the healthcare system, or the time-period. Hence this external dataset may need to be adjusted, and assumptions shall be required about how the observed data relate to the external dataset.

Parametric analysis of TTE data typically has up to two objectives: to obtain a parsimonious description of the observed data, and/or to predict outcomes for the unobserved future (extrapolation). More work is required into the relative strengths and weaknesses of the alternative models in both settings. For example, for the best-fitting FP model, inclusion of random effects had a negligible impact on the AIC. Further research is required to see if this is a general phenomenon, or if more nuanced modelling would lead to a more substantive improvement in fit, or that these enhancements would be beneficial for other observed hazard patterns. The case-study also highlights that a within-sample measure of GoF cannot be used to choose between models for extrapolation, as has been observed previously [23, 7, 22]. The case-study expands on these findings as it compares global models (FPs and survival models), piecewise models (spline-based models), and local models (DSMs). Further work on model choice when used for extrapolation could build upon the work of forecasting competitions [250].

The case-study had limitations. First, we compared models based on AIC (within-sample) and SSE (extrapolations). We were not able to test the differences for statistical significance. For AIC, there is some guidance on what differences may be important, but this only holds for nested models [201]. Whilst the more flexible models generally improved within-sample fit, they did not improve extrapolation performance. In addition, for many analysts, use of the more flexible models will come at an additional ‘cost’ as there will be a need to understand both the theoretical details (strengths and limitations) of the method, as well as how to implement the model. The guidance of Section 3.4 and the reproducible case-study should help to reduce these costs, although they will still be a factor when choosing between the difference models.

The use of a single case-study may also be viewed as a limitation. It is unclear if the (generally) superior GoF provided by DSMs and GAMs generalises to other settings. The results for the three DSMs illustrate an important caution against generalisation: if only the two DSMs without a local trend were considered then DSMs would provide the best-fitting models. In contrast, if only the DSM with a local trend were considered than we would conclude that their fit is not as good as spline-based models. The GoF of the DSM with drift also varied markedly between using the full dataset and using the first three years of data. More experience with these different models and their performance for different sample sizes and follow-up times is required before firm conclusions can be made about which (if any) will provide more accurate estimates.

In conclusion, parametric modelling of the hazard function allows for predictions of future outcomes. Standard survival models may be insufficiently flexible to reflect the complexities of observed hazard patterns. The GLM framework and its extension to GAMs, GLMMs and DGLMs can provide insight into the structure of standard one- and two-parameter models, and their assumptions of linearity. In addition to providing more flexible models (as we have demonstrated here), it also allows for a rich class of model specifications via different combinations of the outcome, distribution and response function - although this comes at the cost of needing to understand how and when to implement these models. We have provided guidance to aid with choosing between these models. Further, spline-based GLMs provide a useful alternative to R-P models: with appropriate response function these models cannot estimate implausible negative hazards, unlike R-P models. A motivating and fully reproducible case-study has demonstrated that these currently under-used approaches can sometimes provide better GoF and more plausible extrapolations than standard survival models.

Appendix 4

Appendices: Generalised linear models for flexible parametric modelling of the hazard function

4.1 Flexible generalised linear models

This section describes flexible models that may be applied within a GLM framework, and hence may be used for the analysis of survival data. To aid interpretation, the focus is on situations where the only covariate of interest is time (a common occurrence in HTA), although extensions to additional covariates are straight-forward. Subsequent sections describe extensions to GLMs.

4.1.1 Fractional polynomials

FPs have been developed to provide a systematic framework for identifying and modelling non-linear effects of a continuous variable [133]. The degree of flexibility of an FP is defined by its order (how many terms it includes). For a single variable, an i^{th} order FP is defined as:

$$E(y_t) = \text{FP}(i) = \beta_0 + \sum_{j=1}^{i} \beta_j x^{p_j} \tag{4.1}$$

where typically the powers p_j are chosen from the set {-2, -1, -0.5, 0, 0.5, 1, 2, 3}, with x^0 denoting $\log x$. If the power of a term is duplicated then the duplicated term is multiplied by $\log x$. Hence, for example, a second order FP of a single variable may be written as $\text{FP}(2) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$ if $p_1 \neq p_2$ and $\text{FP}(2) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} \log x$ if $p_1 = p_2$. It is possible to consider FPs with order > 2, and values of p_j other than those described. However, in practice these extensions are unlikely to lead to improvements in goodness of fit [251]. Similarly, other possible values for p_j may be considered, but may not. For any given order, the powers to use may be based on minimising the AIC. To choose the order of FP, the following (approximate) closed test procedure may be used [133]:

- Overall association of the outcome with time, comparing FP(2) model with model omitting time (Non-significant result = stop testing, do not include time in model).
- Evidence for non-linearity, comparing FP(2) model with a model that is linear in time (Non-significant result = stop testing, use a linear model: power = 1).
- Simpler or more complex non-linear model, comparing FP(2) model with FP(1) model (Significant result = use FP(2) model, non-significant = use FP(1) model).

Comparisons are performed using likelihood ratio tests, with a pre-specified significance level. A limitation of FPs is that, when using the eight powers described above, the continuous variable has to be > 0. This is not a restriction for TTE data, but it can be a problem if transformations of time are used. For example, the generalised F (and hence its special cases, which include the generalised Gamma, Weibull, lognormal and log-logistic amongst others[130]) use the logarithm of time, which can be negative. Of the eight powers described above, only five may be used with negative values: {-2, -1, 1, 2, 3}. Further, it is not possible to handle repeated powers, which would result in taking the lograithm of the covariate. Hence, for log time there are five FP(1) models and 10 FP(2) models. For time as a covariate (no log transformation) there are eight FP(1) models and 36 FP(2) models. If the closed test procedure rejects a non-linear model, and a Poisson GLM is used, then the chosen model will be the same as a standard TTE model. A linear model in time is analogous to a Gompertz model, a linear model in log-time is analogous to a Weibull model, and a model without time as a covariate (and so just an intercept) is analogous to an exponential.

Limitations

Whilst FPs are flexible and relatively parsimonious, they have some limitations:

- Insufficient power to detect non-linearity [133]
- Inability to model a variety of functional forms including logarithmic functions and 'threshold effects' [252]
- Lack of invariance with respect to the coding used for covariates: any transformations such as centering or scaling can lead to different FPs being chosen [253]
- Reduced options for modelling covariates that can be positive.
- Sensitivity to extreme values in the data [207]

The last limitation can be a particular issue if extreme values occur near the end of follow-up time, with subsequent implications for extrapolation.

4.1.2 Restricted cubic splines and Royston-Parmar models

A cubic spline represents a continuous function as a series of piecewise cubic polynomials [124]. These cubic polynomials are restricted to join (have the same value) at a set of 'knots'. The complexity of the cubic polynomial representation depends on how many knots are used. There will always be at least two knots, known as 'boundary' knots. Any additional knots are known as 'interior' knots and occur between the boundary knots. Hence for k interior knots the knot locations are $\xi_{\min} < \xi_1 < \cdots < \xi_k < \xi_{\max}$, with the boundary knots being $\{\xi_{\min}, \xi_{\max}\}$. Cubic splines also have continuous first and second derivatives.

A RCS is further restricted to be linear beyond the boundary knots (that is, before $\xi_{\min}$ and after $\xi_{\max}$) [136]. A spline has several alternative formulations. One common description for an

RCS is as a truncated power series [136, 252]:

$$
\begin{aligned}
E(y_t) &= \beta_0 + \beta_1 x + \beta_2 V_1(x) + \cdots + \beta_{k+1} V_k(x) \\
\text{with } V_j(x) &= (x - k_j)_+^3 - \psi_j (x - k_{\min})_+^3 - (1 - \psi_j)(x - k_{\max})_+^3, \quad j \in \{1, \ldots, k\} \\
\text{and } \psi_j &= (k_{\max} - k_j)/(k_{\max} - k_{\min}) \\
\text{where } (x - a)_+ &= \max(0, x - a)
\end{aligned}
\tag{4.2}
$$

where the $V_j(x)$ are referred to as basis functions. Alternative bases are formed by b-splines or p-splines (which are a function of b-splines). B-splines are numerically more stable than the truncated power series basis. However, in practice this advantage is usually negligible, and b-splines cannot be used to generate extrapolations, whereas the truncated power series basis can [252].

In the absence of any interior knots, an RCS is a linear function. Hence, as previously noted, standard survival models may be obtained as special cases of RCSs. For extrapolations beyond $k_{\max}$ the RCS is a linear function of x.

For a Poisson GLM, the outcome is the hazard rate. An alternative application of RCSs, outside the GLM framework, is with an R-P model, for which the log cumulative hazard is the outcome [136]. This is motivated by noting that the hazard function may be more noisy than the cumulative hazard function. This increased noise is most likely near the end of follow-up time, due to small numbers. These 'end effects' may induce spurious artefacts in the spline function. A drawback of directly modelling the cumulative hazard is that this is monotonic (it cannot decrease), but this property cannot be expressed by a simple set of constraints on the parameters of the spline. Hence it is possible that implausible functions may be fitted, as a modelled decreasing cumulative hazard would imply that there is a period for which the hazard rate is negative. An example of implausible fit to real-data is provided in Figure 4b of [134]. Further developments of the R-P model are described by Lambert and Royston [254, 124].

The model specification for a RCS includes both the number of internal knots, and also their placement. It has been suggested that the former decision is more important than the latter and that a maximum of three internal knots is usually sufficient, as larger values may lead to overfitting the data [136, 252]. When introducing the R-P model, the authors also suggested that knots be placed at equally-spaced intervals of the uncensored event times (with the boundary knots placed at the first and last of the observed uncensored event times)[136]. The choice of how many interior knots to use is typically based on minimising an information criteria.

Limitations

The main limitation with RCS is that, due to their flexibility, they may over-fit local 'noise' in the data [205]. The can also be 'data hungry', as they can produce biased results in small samples [255]. In addition, unlike FPs, there is in general no closed test procedure for choosing between different models. One exception is that linear models are always nested within more complex models, so a test for non-linearity is possible.

4.2 Generalised linear mixed models

The presence of unmeasured variables can lead to heterogeneity (overdispersion) in the outcome that is in excess of what is implied by a model. This will lead to underestimating uncertainty in covariate effects [107]. To remedy this, GLMs may be extended by incorporating random effect terms, resulting in GLMMs. These terms have a mean of zero and a variance parameter, usually unknown. Frailty refers to the feature that individuals have unequal conditional probabilities of experiencing the outcome of interest, given these random (frailty) terms. [17]. Frailty models are appropriate when data are collected on multiple levels; for example individuals may be clustered within families, or they may be clustered by centre within a multicentre clinical trial. Because of this, frailty models may also be referred to as multi-level models [91].

Frailty models may be used to analyse TTE data outside a GLMM framework [107]. However, with the GMLM framework, frailty models may be modelled as mixed models, which have been extensively researched [256, 246]. Mixed models are so-called as they contain both fixed effects and random effects. Fixed effects occur in a standard GLM; for these effects (parameter coefficients) are assumed to apply (are fixed) for all individuals. In contrast, random effects are allowed to vary across individuals. To avoid identifiability issues, it may be assumed that the random effects are drawn from a zero-mean Normal distribution (other zero-mean distributions such as t-distribution may also be used).

As an example, consider a fixed-effects 2nd-order FP which, if powers are not duplicated, may be written as:

$$E(y_t) = \text{FP}(2) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2}$$

When analysing TTE data with a GLM, individual observations relate to individual time intervals. As there is only one observation per time-interval, a GLMM with a random intercept may be used. That is, the linear predictor in any time-interval t is increased or decreased by an additive amount b_t. For example:

$$E(y_t) = \text{FP}(2) = \beta_0 + b_t + \beta_1 x^{p_1} + \beta_2 x^{p_2}, \quad b_t \sim N(0, \psi^2) \tag{4.3}$$

The above extension is not specific to FP models. That is, any fixed-effects GLM may be extended by adding a random intercept as shown above.

4.2.1 Limitations

As the frailty terms are not observed, it may not be possible to identify them from the data, which can lead to problems with model specification and estimation. [107, 139]. A further limitation with the use of frailty models is that it is unclear how random effect terms should be extrapolated. One option is to fit a model to the estimated random effects and use this to predict future values. Alternatively the last estimated random effect may be carried-forwards.

4.3 Generalised additive models

A generalised additive model (GAM) may be viewed as a GLM in which a covariate x is replaced by a linear combination of a finite set of smooth functions $b_1, \ldots, b_q$ [248]. For example, a univariate GAM is defined as:

$$E(y_t) = \sum_{j=i}^{q} b_j(t)\beta_j = f(t) \tag{4.4}$$

Where $b_j(t)$ is the jth basis function, and q is the dimension of the basis function. Modelling the effect of a covariate as a sum of basis functions results in extremely flexible models, with more flexible models arising as more functions are added to the basis. The basis functions may be non-parametric, which results in semi-parametric GAMs [248]. However, the focus here is on parametric basis functions, and hence parametric GAMs [135].

Of the parametric basis functions, use of RCSs is of particular interest. Model estimation for a GAM is different to that for a GLM. For the latter, model complexity may be penalised after model fitting via the use of information criteria, such as the AIC. For a GAM, model complexity is included in the objective function to be minimised. For example, for Normally distributed data this would be [135]:

$$\sum_{t=1}^{n}\Big(y_t - f(t)\Big)^2 + \Lambda \int \Big(f''(t)\Big)^2 dt \tag{4.5}$$

The integral for the second term is taken over the range of the data and penalises the wiggliness of function, to avoid over-fitting the data. The first term in equation (4.5) quantifies model goodness of fit. In this example it is the sum of squared errors, but a more general representation uses a negative log-likelihood, as is used in a GLM. This leads to the following objective function [247]:

$$-2\sum_{t=1}^{n}\mathcal{L}_t\Big(y_t; f(t)\Big) + \Lambda \int_{-\infty}^{\infty} \Big(f''(t)\Big)^2 dt \tag{4.6}$$

As with AIC, the objective function measures the trade-off between model goodness of fit and model complexity. The key difference is that for GAMs the complexity of the model is estimated during the model-fitting process, and is not a post-hoc choice. For a GAM, Λ quantifies this trade-off. As $\Lambda \to \infty$ the estimated function tends towards a straight line. For $\Lambda = 0$ there is no penalty on the regression spline (so the resulting model will be equivalent to a RCS model if a RCS basis is used).

There are two components to model estimation, as the smoothing parameter Λ and the model coefficients β are estimated separately [135]. One approach to estimating the degree of smoothing is based on a generalisation of leave-one-out cross-validation, known as generalised cross-validation. For given values of Λ, model coefficients are estimated based on penalised likelihood estimation [257].

As with standard GLM models, GAMs may also be extended to incorporate random effects. The resulting models are known as generalised additive mixed models [258]. This extension does not result in any concepts additional to those outlined in section 4.2.

4.3.1 Choice of basis function

A full RCS (with a knot at each unique value of t) has the desirable property of being a 'smoothest interpolator' [135]. That is, for all functions that are continuous over the range of the data and also have an absolutely continuous first derivative, a full RCS will minimise equation (4.6) [247]. In addition, a RCS is usually flexible enough to adequately describe most functions. An alternative choice of basis function is to use polynomials. Here the j^{th} basis function is defined as [135]:

$$b_j(t) = t^{j-1} \tag{4.7}$$

In comparison with equation (4.2), the basis for polynomials is much simpler. The first two terms are also the same as for a RCS, and so the use of polynomial basis functions also includes standard models such as the Weibull and Gompertz as special cases. However, the use of polynomial basis functions is not recommended [135, 252]. Unlike RCSs, there is no guarantee that polynomials will provide an adequate fit across the entire range of the data. Further, the fit in a local region of the data can be strongly affected by the characteristics of the data in other regions. Also, polynomials are unable to provide adequate descriptions of a number of functions, such as those including 'threshold effects' or logarithmic functions. It should be noted that FPs (section 4.1.1) do not follow the hierarchical rule that is used when constructing polynomial basis functions. Indeed, FPs include fractional powers, which are not in the definition of a polynomial basis function in equation (4.7). Hence, unlike R-P models, FPs cannot be viewed as a special case of a GAM. As such the limitations that apply to the use of polynomial basis functions do not necessarily apply to the use of FPs.

4.3.2 Regression splines

A full smoothing spline has as many parameters as there are data points. These full splines may be approximated by regression splines, which are no longer 'smoothest interpolators', but are computationally easier to fit [135]. A simulation study comparing full splines with their regression approximation found that the approximation led to a superior fit (to the known 'true' function) [259]. The author attributed this counter-intuitive finding to the fact that regression splines are also less likely to overfit the data (or equivalently, the full splines was modelling random variation).

When using regression splines, the basis dimension has to be specified (it is not part of model estimation). In practice the smoothing parameter Λ in equation (4.6) has a larger impact on model complexity than the choice of basis dimension. Hence the basis dimension acts as an upper-bound for model complexity. Provided it is sufficiently large, model results are typically insensitive to the choice of basis dimension [135]. A slight exception is that larger basis dimensions lead to a larger set of candidate models, which can sometimes affect model choice.

4.3.3 Limitations

When using RCSs as basis functions, GAMs share the same limitation in that they require sufficiently large datasets, and they may over-fit local noise in the data [259]. However, the use of a penalized objective function slightly mitigates this latter limitation.

4.4 Dynamic generalised linear models and dynamic survival models

A DGLM extends a GLM by allowing the model coefficients to vary as a smooth function of time. These may be used to analyse TTE data, resulting models are known as dynamic survival models (DSMs). As an illustrative example, consider the exponential model: $\log(\mu_t) = \beta_0$. This may be interpreted as a global level model as the outcome is set to a fixed level for all times. This assumption can be relaxed by allowing β_0 to vary over time. Hence $\log(\mu_t) = \beta_{0,t}$, giving a local-level model.

Without further restrictions, this may lead to an over-fitting model. For example, setting $\beta_{0,t} = y_t$ would give a perfect fit to the observed data. However, this will over-fit the data, so to avoid this the following restriction is used:

- When estimating the local level at time t, only evidence available prior to time t is used. This can include previous estimates of the local level, previous covariate values, and previous observations.

More formally, at time t, let the prior available evidence be denoted by the history of prior outcomes: $\mathcal{H}(y)_t = \{y_1, y_2, \ldots, y_{t-1}\}$, the history of prior coefficient estimates $\mathcal{H}(\beta)_t = \{\beta_1, \beta_2, \ldots, \beta_{t-1}\}$, and the history of prior covariates $\mathcal{H}(x)_t = \{x_1, x_2, \ldots, x_{t-1}\}$. Then the estimate of the local level at time t is obtained as a function of all these histories $\mathcal{H}(y, \beta, x)_t, = \mathcal{H}_t$ for simplicity. This estimate is a forecast (extrapolation) from the evidence at time $t-1$ to time t. As such, it is referred to as a *one-step-ahead forecast*.

The approach of DSMs represents a small, but significant change in estimation from previously described approaches. For these, a model estimate of y_t (denoted by $\hat{y}_t$) used all the observed data. Such an estimate is known as a *smoothed estimate*, and models minimised the error of smoothed estimates. In contrast, DSMs minimise the error of (within-sample) forecasts. Hence the choice between a DSM and a GLM (or GAM) is likely to be based on if the analyst wants to minimise the error of smoothed estimates or the error of forecasts.

Some further comments on the use of the forecasts are required.

- It shall be assumed that the Markov property is valid [178]; so $\hat{y}_{t|1:t-1} = \hat{y}_{t|t-1}$. In addition, sometimes an estimate at time t uses y_t. This is no longer a one-step ahead forecast, but is referred to as a *filtered estimate*, denoted as $\hat{y}_{t|1:t} = \hat{y}_{t|t}$.

- When estimating the one-step ahead forecast $\hat{y}_{t+1|t}$, evidence is available for all the prior one-step ahead forecasts (which are a function of the β_t). However, evidence is also available on the previous outcomes. This may be used to estimate the accuracy of the previous forecasts, and this knowledge used to improve estimates of $\hat{y}_{t+1|t}$. Hence the one-step ahead forecast $\hat{y}_{t+1|t}$ is based on prior filtered values, and so the histories $\mathcal{H}(\beta)_t$ include filtered values, not one-step ahead forecasts.

- Specification of a DSM requires initial (starting) estimates. There will be no histories to inform these, so they may be based on external data, expert opinion, or via a heuristic [42].

A limitation with DSMs is that model estimation can be difficult: either due to computational issues, or if the use of approximate methods is inappropriate [168, 144]. Extrapolations from a DSM are determined by the chosen model specification, which is described in the following subsections.

4.4.1 Model specification

The general model specification for a DGLM (or DSM) is [247]:

$$\text{Observation model: } E[y_t] = \mu_t \quad y_t \sim \text{exponential family distribution} \tag{4.8a}$$

$$\text{Response function: } \mu_t = h(x_t^T \beta_t) \tag{4.8b}$$

$$\text{Transition model: } \beta_t = F\beta_{t-1} + \zeta_t \tag{4.8c}$$

$$\text{Initial conditions: } \beta_0 \sim MVN(b_0, Z_0) \tag{4.8d}$$

where MVN denotes a multivariate Normal distribution, F is a transition matrix for the coefficients over time, and the error term (ζ_t), also referred to as an *innovation*, is assumed to be an independent and identically distributed series with $\zeta_t \sim N(0, Z_t)$, where Z_t is a variance-covariance matrix of the model coefficients. As before, for a Poisson GLM, the outcome is the number of number of events in an interval (which, combined with knowledge of the at-risk population, can be used to derive the hazard rate). Of note, the covariate time does not appear within the model specification. That is, no assumptions are made about the relationship between time and the hazard rate. Instead it is assumed that there exist latent states (the βs), such as an average value (level) and trend - either of which may vary over time.

Three types of DLM are of particular interest. Two are the local level and local trend models, which may be interpreted as zero-order and first-order Taylor series approximations, respectively [42]. The third is the local level global trend model, also referred to as a local level with drift model. These are defined below.

Local level models

$$
\begin{aligned}
E(y_t) &= \beta_t \\
\beta_t &= \beta_{t-1} + \zeta_t
\end{aligned}
$$

this type of model is referred to as a random-walk model. Extrapolations of the y_t are a constant value, equal to the last-estimated local level. That is, the extrapolated value h time steps into the future, given observed data up to time T is $\hat{y}_{T+h|T} = \beta_T$,

Local trend (linear) models

These may be defined as:

$$
\begin{aligned}
E(y_t) &= x^T\beta_t = [1,0]\begin{bmatrix}\beta_{1t}\\ \beta_{2t}\end{bmatrix} = \beta_{1t} \\
\beta_t &= \begin{bmatrix}\beta_{1t}\\ \beta_{2t}\end{bmatrix} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}\beta_{1,t-1}\\ \beta_{2,t-1}\end{bmatrix} + \begin{bmatrix}\zeta_{1t}\\ \zeta_{2t}\end{bmatrix} \\
\beta_{1t} &= \beta_{1,t-1} + \beta_{2,t-1} + \zeta_{1,t} \\
\beta_{2t} &= \beta_{2,t-1} + \zeta_{2,t}
\end{aligned}
$$

where $\beta_t = [\beta_{1t}, \beta_{2t}]^T, \zeta_t = [\zeta_{1t}, \zeta_{2t}]^T$. The formula for β_{1t} may be written down recursively as:

$$
\begin{aligned}
\beta_{1t} &= \beta_{1,t-1} + \beta_{2,t-1} + \zeta_{1,t} \\
&= \beta_{1,t-2} + 2\beta_{2,t-2} + \zeta_{1,t} + \zeta_{1,t-1} + \zeta_{2,t-1} \\
&= \beta_{1,t-3} + 3\beta_{2,t-3} + \zeta_{1,t} + \zeta_{1,t-1} + \zeta_{1,t-2} + \zeta_{2,t-1} + 2\zeta_{2,t-2} \\
&= \dots \\
&= \beta_{1,0} + t\beta_{2,0} + \sum_{i=1}^{t}\zeta_{1i} + \sum_{j=1}^{t-1} j\zeta_{2,t-j}
\end{aligned}
$$

where the first two terms are a linear function, and the last 2 terms are random error. As such, β_{1t} represents the local level, which is a linear function of t. The parameter β_{2t} may be interpreted as the growth rate (local trend). Extrapolations are a linear function of the last-estimated local level and local trend values: $\hat{y}_{T+h|T} = \beta_{1T} + h\beta_{2T}$.

Local level global trend (drift) models

This may be viewed as a special case of the local trend model, with $\zeta_{2,t} = 0$. A single trend β_2 is then estimated based on all the data. This estimate of the trend will be more stable than the one estimated locally as towards the end of follow-up the number of observations is smallest. However, this increased stability is at the loss of flexibility, as the estimate of trend is fixed and so does not vary with time.

4.4.2 Likelihood specification

The likelihood for a DSM may be written as the product of likelihoods for individual time-periods j (denoted by $\mathcal{L}_j$), with t_j denoting the jth time-period. Following Hemming and Shaw [40] the following two variables are introduced:

$$\delta_{ij} = \begin{cases} \delta_i & \text{if } t_i^* \leq t_j \\ 0 & \text{if } t_i^* > t_j \end{cases} \qquad t_{ij} = \begin{cases} t_i & \text{if } t_i^* \leq t_j \\ t_j & \text{if } t_i^* > t_j \end{cases} \tag{4.9}$$

hence for an individual who experienced an event, $\delta_{ij} = 1$ for all time periods up-to and including the time period which included the event. In all other situations it is zero. The indicator t_{ij} is the observed survival time up-to and including the time period which included the observation. Subsequently it is set equal to the time interval. The likelihood for a Poisson DSM is then:

$$\mathcal{L} = \prod_{j=i}^{N} \mathcal{L}_j = \prod_{j=i}^{N} \prod_{i=i}^{\tau_j} \exp(-[t_{ij} - t_{j-1}]e^{x_t^T \beta_j}) e^{x_t^T \beta_j \delta_{ij}} \tag{4.10}$$

where N is the number of time-intervals, and τ_j is the set of individuals who have a survival time $\geq t_j$. When equal interval widths are used, this is the same likelihood as for a Poisson GLM with an offset (see Section 4.5).

4.4.3 Limitations

The flexibility of allowing model coefficients to vary as a function of time can lead to problems with both convergence, and specification of initial estimates [128, 143, 40, 144].

4.5 Likelihood estimation

4.5.1 Parametric survival models

Let d_i be an event indicator which $= 1$ if t_i is an observed survival time and $= 0$ if t_i is a right-censored observation (hence the true survival time will be greater than t_i). The likelihood is then [100]:

$$\mathcal{L}_i = \left\{ \prod_{i:\ d_i=1} f_i(t_i) \prod_{i:\ d_i=0} S_i(t_i) \right\} \tag{4.11}$$

For individuals with an observed event their probability density function contributes to the likelihood. Individuals with censored times contribute their cumulative survivor function (as their probability density function is not fully observed but it is known that they survived up to time t_i). The log-likelihood contribution for the i^th patient is:

$$\begin{aligned} \log \mathcal{L}_i &= \log \left\{ f(t_i)^{d_i} S(t_i)^{1-d_i} \right\} \\ &= d_i \log\{f(t_i)\} + (1-d_i)\log\{S(t_i)\} \end{aligned} \tag{4.12}$$

The following identities may be used to derive an expression of the likelihood purely in terms of the hazard function [98]: $S(t) = 1 - \int_0^t f(u)du$ and $f(t) = h(t) \times S(t)$. This gives:

$$\begin{aligned} \log \mathcal{L}_i &= \log \left\{ h(t_i)^{d_i} S(t_i) \right\} \\ &= d_i \log\{h(t_i)\} - \int_0^{t_i} h(u)du \end{aligned} \tag{4.13}$$

4.5.2 Equivalence between a Poisson GLM with an offset and a Poisson DSM

Poisson GLM with an offset.

The probability density function is:

$$f(y_i; \tau_i\theta) = \frac{\tau_i\theta^{y_i} e^{-\tau_i\theta}}{y_i!}$$

The likelihood $\mathcal{L}$ is equal to the sum of the y_i. Taking logarithms and re-arranging gives:

$$\begin{aligned} \mathcal{L} &= \prod \left(\frac{\tau_i\theta^{y_i} e^{-\tau_i\theta}}{y_i!} \right) \\ \log \mathcal{L} &= \sum \left[\log \left(\frac{\tau_i\theta^{y_i} e^{-\tau_i\theta}}{y_i!} \right) \right] \\ &= \sum \left[\log(\tau_i\theta^{y_i}) + \log(e^{-\tau_i\theta}) - \log(y_i!) \right] \\ &= \sum \left[y_i \log(\tau_i\theta) - \tau_i\theta - \log(y_i!) \right] \\ &= \sum \left[y_i \log(\tau_i) + y_i \log(\theta) - \tau_i\theta - \log(y_i!) \right] \end{aligned}$$

The first and last terms do not include the parameter θ. Thus, for a given dataset, the first and last terms will always be the same, and so may be eliminated from the likelihood when comparing different models. This gives the likelihood:

$$\log \mathcal{L} = \sum \left[y_i \log(\theta) - \tau_i\theta \right] \tag{4.14}$$

Poisson DSM.

Consider the Poisson DSM likelihood for any given time-period.

$$\mathcal{L}_j = \prod_{i=i}^{\tau_j} \exp(-[t_{ij} - t_{j-1}]e^{z\beta_j})e^{z\beta_j\delta_{ij}}$$

Assume that intervals are all of the same width, so that the term $[t_{ij} - t_{j-1}] = 1$ and may be omitted. Taking logarithms:

$$\begin{aligned}\log \mathcal{L}_j &= \sum_{i=i}^{\tau_j} \log\left[\exp(-e^{z\beta_j})e^{z\beta_j\delta_{ij}}\right] \\ &= \sum_{i=i}^{\tau_j} \log[\exp(-e^{z\beta_j})] + \log(e^{z\beta_j\delta_{ij}}) \\ &= \sum_{i=i}^{\tau_j} -e^{z\beta_j} + z\beta_j\delta_{ij} \\ &= d_j z\beta_j - \tau_j e^{z\beta_j}\end{aligned}$$

where d_j is the observed number of events in the jth time-interval, hence $= y_j$. Let $\theta_j = e^{z\beta_j}$. Then:

$$\log \mathcal{L}_j = y_j \log(\theta_j) - \tau_j\theta_j \tag{4.15}$$

which is the same as applying equation (4.14) to a single time-period.