Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Statistical Methods to Extrapolate Time-To-Event Data

Philip Cooney

February 24, 2024

Submitted in partial fulfilment of the requirements for the doctorate of philosophy in statistics.

PhD School of Computer Science and Statistics

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

I agree that this thesis will not be publicly available, but will be available to TCD staff and students in the University's open access institutional repository on the Trinity domain only, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: Philip Cooney   Date: February 24, 2024

Abstract

This thesis investigates methods used to predict long-term survival of observations (typically survival times) beyond the time at which data follow-up is available. Current practice is to use parametric survival models; however, different models can produce different survival predictions, particularly if the lifetimes of many of the observations are censored.

We focus on applying novel statistical techniques to improve existing methods to predict survival. One existing predictive approach assumes that after a certain timepoint, the hazards are approximately constant, and a constant hazard after this timepoint is used to estimate long-term survival. The choice of this timepoint is arbitrary and subject to considerable uncertainty. To improve on this methodology we estimate a statistical model known as a change-point survival model. This model allows the observed data to inform the timepoint after which the constant hazard is appropriate. Statistical goodness of fit measures can identify if the addition complexity associated with the inclusion of a change-point is warranted. We also estimate other more complex change-point survival models which allow us to model multiple treatments.

Another topic which was investigated is the incorporation of expert opinion with statistical models. In the case of survival predictions, even if the survival is not observed at a timepoint, there are often opinions on the plausible ranges that these values may take. In this thesis, we investigate how these opinions can be incorporated in a robust manner, allowing for the predicted survival to take account of the precision of the expert's opinion and the sample size of the observed data. We also estimate how to quantify the strength of an expert's opinion to allow for appropriate calibration of their opinions at the elicitation stage.

We found that the change-point model we estimated can robustly detect the timepoint at which a constant hazard is appropriate. In several real-world applications, it provided the closest predictions to the follow-up survival data. The proposed method for incorporation of expert opinion allowed for the straightforward synthesis of different types of expert opinions with data. We demonstrate by way of a simulation study that including expert opinion can more accurate survival predictions, even when the expert's belief is biased away from the true estimate. By numerically quantifying the strength of expert's beliefs, we more easily identify situations where expert's opinions are overconfident, allowing for re-calibration of their beliefs.

The key methods from the thesis are implemented as open-source software packages to allow the methods to be used in practical applications. The ideas in this thesis can also be extended and improved upon in future research. We believe that the methods illustrated in this work will improve the ability of decision makers to model hypotheses relating to the prediction of long-term survival outcomes.

Lay Abstract

In this research, we aim to improve the prediction of survival outcomes by applying a variety of novel statistical techniques. Often, information on survival outcomes is incomplete with observations (or patients) still alive at the end of a study. To obtain long-term survival predictions we can use one of several statistical models. The challenge is to identify which statistical model is most appropriate, especially since different models can provide quite different predictions.

One technique is to assume that after a certain timepoint, the annual probability of surviving is the same for every time interval. That is, the probability of surviving from 2 years to 3 years is the same as surviving from 5 years to 6 years. We estimate a statistical model that allows us to determine this timepoint from the data rather than requiring it is chosen manually. In subsequent research, we estimate more complex models which allow us to determine the timepoint after which each treatment is equally effective.

Another topic investigated is the incorporation of expert opinion with statistical models. Experts may have a variety of beliefs that they would like the statistical model to reflect. We develop methods which allow for these beliefs to be incorporated with data, allowing for a final model which is a true reflection of what the expert believes and what is observed from the data.

By applying our methods to real world examples and computer simulations we find that our approaches provide more robust predictions of long-term survival. We have produced software programmes which allow for the widespread use of our methods by non-experts, therefore, enhancing the impact of this research. This thesis poses several new avenues for further research in the topic of predicting survival times.

Acknowledgements

Firstly I would like to thank my family who have always supported me and driven me to achieve my academic and personal goals. A special thank you to my partner Giovanna who has been with me for the majority of my PhD and whose love and support means everything to me. My education has allowed me to meet and become friends with many wonderful people, both at secondary school and the universities which I have attended. These people have made me the person I am today and are more valuable to me than any degree or academic publication. My thanks also go to my supervisor Arthur, who throughout my PhD journey has always provided encouragement and support for the research we conducted. As I was doing this PhD I was also an employee of Novartis and none of this research could have happened without the support I received from my long-time manager (and friend) Ronan Mahon. Finally I would like to thank all the users of statistical, programming, Latex forums who spend time and effort to answer questions and those who create open source software. Without their contributions, academic and professional progress would certainly be hampered if not completely stopped. Throughout history many scholars have noted "If I have seen further it is by standing on the shoulders of Giants" and this is certainly true for the contribution provided by this research.

Contents

List of Figures

List of Tables

Nomenclature

Notation which is frequently used within the thesis is presented below, along with all acronyms.

t	time
T	a random variable of a survival time
t_i	an observed time for individual
$\mathbf{t}_{1:n}$	Time ordered vector of survival times
$f(t)$	probability density function; typically for time t but generically for any other variable
$F(t)$	Cumulative distribution function for time t
$S(t)$	Survival function for time t
$h(t)$	hazard function for time t
$H(t)$	Cumulative hazard function for time t
D	Observed data
ν_i	Individual's indicator function for event/censoring time
θ	Generic notation for model parameter (boldface if vector)
$p(\theta)$	Prior for parameter
$L(\theta\|D)$	Likelihood of parameter
$\pi(\theta\|D)$	Posterior distribution of the parameter
$\pi(D\|\xi)$	Marginal likelihood (possibly conditional on hyperpriors/hyperparameter ξ)
$\tau_{1:k}$	Vector of k change-point times
$I(x > y)$	If condition is true the function returns 1 and 0 otherwise
$Q(p)$	Quantile function
$\mathcal{G}$	Gamma distribution
$\mathcal{IG}$	Inverse Gamma distribution
$\mathcal{NG}$	Normal Gamma distribution
$\mathcal{B}$	Beta distribution
$\mathcal{E}xp$	Exponential distribution
$\mathcal{U}$	Uniform distribution
$\mathcal{LN}$	Log-Normal distribution
W	Weibull distribution
$\Gamma(x)$	Gamma function
$\Gamma(x, s)$	Upper incomplete gamma function

Abbreviations

AIC	Akaike Information Criterion
ATF	Accelerated Time Factor
B&B	Bagust and Beale (approach)
BIC	Bayesian Information Criterion
CEA	Cost-Effectiveness Analysis
DIC	Deviance Information Criterion
ESS	Effective Sample size
GPM	General Population Mortality
HMC	Hamiltonian Monte Carlo
HTA	Health Technology Assessment
LAP	Loss adjusted posterior
LKJ	Lewandowski-Kurowicka-Joe (distribution)
MAP	Maximum a Posteriori
MCMC	Markov Chain Monte Carlo
MLV	Most Likely Value
MNAR	Missing Not at Random
NICE	National Institute for Health Care Excellence
PH	Proportional Hazards
PML	Psudeo-Marginal Likelihood
PSRF	Potential Scale Reduction Factor
QALY	Quality Adjusted Life Years
RJMCMC	Reversible Jump MCMC
SHELF	Sheffield Elicitation Framework
TA	Technology Appraisal
TSD	Technical Support Document
WAIC	Widely Applicable Information Criterion

Part I

Part Introduction and Overview

1 Introduction

The primary aim of many studies is to analyse the time until a pre-specified event of interest occurs. In these settings, the response variable is the time until that event, which is often called failure time, survival time, or event time. Time-to-event data are usually not observed for all observations under study, primarily because the data from a study are analysed at a point in time when some individuals are still alive, resulting in these observations being censored.

In clinical trials that have time-to-event outcomes, the primary objective is to identify if there is a statistically significant difference in the expected survival times of the treatment arms. In other disciplines such as health economics, the primary focus is to assess the long-term expected survival of both treatment groups so that the incremental health outcome of an intervention can be calculated. Except in situations where we are willing to assume that the long-term difference in health outcomes is similar to the differences in survival observed in the trial (see Monnickendam et al. (2019)), we are required to assume a parametric model for the data generating process. These parametric models provide survival functions to predict (or extrapolate) the long-term survival and calculate the average time survived.

Any valid probability distribution that has a support from $[0, \infty)$ can in principle be used for this purpose, and each distribution implies a particular functional form of the hazard function. From the exponential distribution, which assumes a constant hazard, to the four-parameter generalized-F distribution which can accommodate bathtub type hazards, the choice of model will determine the hazard function and consequently the expected survival. Differences in long-term predictions can be particularly pronounced when a high proportion of survival times are censored and may produce clinically implausible survival estimates. This issue has been discussed by Davies et al. (2013) among others, and several solutions have been proposed, including model averaging (Jackson et al., 2010), using external data (Guyot et al., 2017), and expert opinion (Cope et al., 2019). In this thesis, the primary focus is the prediction of long-term survival using novel statistical methods in the decision modelling of health interventions, which we will refer to as Health Technology Assessment (HTA). It should be noted that the methods described in this thesis could also be considered in the field of reliability analysis and, in particular, the work on expert opinion can be applied to almost any area employing statistical models.

1.1 Overview and Motivation

Various stakeholders such as clinicians, patient groups and pharmaceutical companies, are interested in the long-term survival associated with various treatment options, however, this is arguably most important for decision makers who decide which treatments should be made available through publicly funded health services.

Consequently, decision makers such as NICE provide guidelines for HTA in general and more specifically regarding the extrapolation of long-term survival outcomes. Other guidelines are provided by special interest groups such as ISPOR (The Professional Society for Health Economics and Outcomes Research) and in publications in journals with a focus on HTA. The primary guidance documents for extrapolation of long-term survival functions are Latimer (2013) and more recent work by Rutherford et al. (2020) along with an alternative approach considered by Bagust and Beale (2014). Other papers expand upon topics covered in these documents and a more thorough overview is provided in Chapter 2. In the subsequent section we highlight approaches detailed in these documents that can be enhanced using modern statistical techniques.

The first topic which is researched in this thesis relates to estimating change-point models for the hazard function, a particular type of parametric survival model. These models are investigated as a potential alternative to the piecewise models that are sometimes used when extrapolating survival outcomes. Piecewise models are discussed by Rutherford et al. (2020) and defined as having different survival models for each time period, which are typically implemented with an arbitrary number of change-points and without robust justification for the change-point locations. Another potential use case of change-point models is in improving an approach considered by Bagust and Beale (2014), which describes visually inspecting the log-cumulative hazard plot to identify a section of local linearity. This interval is then used to estimate a constant hazard (exponential) model which is used for long-term extrapolation. Both these approaches could be more appropriately estimated using change-point models which estimate both the number and change-point locations directly from the data, allowing full uncertainty in these parameters to be considered rather than fixing them at a subjective location. Estimation of parametric change-point models is discussed in Chapters 3 & 4, with other more complex parametric models considered in Chapter 5.

Another topic which is mentioned in Rutherford et al. (2020) as a research objective is the use of expert opinion to inform extrapolation. This topic forms the second part of the thesis. A general method for including expert opinion on observable outcomes in statistical analyses is developed in Chapter 6, and in Chapter 7 a specific application to survival extrapolation is considered.

Chapter 8 provides an overview of the software packages that have been developed to implement the methods described in this thesis.

Target audience:

The intended audience for this thesis are health economic modellers, who typically possess some level of statistical training. Specifically, we focus on those whose work involves modelling time-to-event outcomes. Throughout the thesis, our goal is to justify why we employ various types of statistical software and overall statistical methodology. This includes discussing choices such as priors and goodness-of-fit statistics. By providing this rationale, practitioners using the methods developed in this thesis will be equipped with the necessary justification to support their modelling decisions when communicating with decision makers and other stakeholders.

1.2 Current problems and scope of present work

Piecewise models are typically employed when there is an apparent lack of fit of survival models to the observed survival data. By partitioning time into a number of intervals and fitting a survival model (including non-parametric estimators) to each of these intervals, a good fit to the observed survival data can be obtained. Manually selecting the locations for these intervals is subjective and artificially reduces the uncertainty associated with the extrapolated survival. By considering fully parametric change-point survival models whose locations are estimated from the data (and potentially informed by subjective opinion) both of the above issues can be addressed. Currently there are few published software programs to estimate this class of problems, and fewer still fully propagate the uncertainty associated with the extrapolated survival as they consider a frequentist framework. Change-point models have the advantage that they can be used to model biologically plausible hypotheses, such as treatment delay or waning which is particularly important when jointly modelling the survival of a treatment and comparator.

The goal of Part II of the thesis is to achieve a method for estimating these models that could meet the following criteria:

1. Data driven selection of the change-points for survival models
2. Fully propagate the uncertainty associated with the model parameters to the extrapolated survival
3. Compare change-point models with a wide class of parametric survival models

which are typically considered in survival extrapolation

4. Develop Software programs which easily output the required outputs for HTA decision modelling (e.g. extrapolated survival)

Another topic which is researched in this thesis is that of expert opinion. Given the intrinsic uncertainty in predicting long-term survival outcomes when expert clinical opinion is available, it is important to use this information in the modelling process. Such opinions are often not integrated in a formal way, for example, survival models are typically estimated using maximum likelihood (i.e., based on the data alone) before choosing the parametric model for which expected survival appears to be compatible with the expert opinion. This approach has a number of weaknesses. Primarily, it is difficult to identify the most appropriate model if several models appear consistent with the expert opinion. In the opposite scenario, when none of the models meet the expert's criteria, the best choice of model is again unclear. It would be preferable to include a measure of statistical fit that takes account of the degree of agreement with the expert opinion as well as the observed data, rather than making a decision based solely on whether the predicted quantity from the model is within the expert's plausible range. Existing methods which incorporate expert opinion with survival models (and statistical models more generally) typically consider only one class of statistical model and one type of expert opinion, rather than a general approach which could easily be applied to a variety of opinion and model types.

Regarding Part III of the thesis we aim to provide a framework for incorporating expert opinion into statistical analysis which is:

1. Widely applicable in terms of the expert opinion and statistical models which can be considered.

2. Implementable with standard commonly used statistical tools without much additional programming.

3. Compatible with the output of commonly used methods for eliciting expert opinion (i.e. SHELF).

The attributes of the method described in Part III align with Mikkola et al. (2023), who specify the criteria for improved techniques relating to incorporating expert opinion.

We connect the findings from Part II to those in Part III, illustrating this with an example that incorporates expert opinion using a change-point survival model.

An overarching objective of this thesis was to make the key contributions available as R-packages which we hope will aid implementation of the methods and allow for further research and improvements of the methods. Chapter 8 is the sole chapter in Part IV and describes in detail the two R-packages, `PiecewiseChangepoint` and `expertsurv` that have been written to implement the piecewise exponential change-point model and incorporate expert opinion with parametric survival models respectively.

1.3 Summary of contributions

In this thesis we propose to estimate a wide variety of change-point survival models using Bayesian framework which allows for estimation and propagation of change-point uncertainty. In particular we consider a change-point model known as the piecewise exponential survival model which we show is an appropriate method to estimate the timepoint after which constant hazards may be considered approximately constant. The identification of this timepoint is a key consideration in the method to extrapolate survival outcomes presented by Bagust and Beale (2014).

We then estimate change-point models for a variety of survival models allowing for the introduction of covariates, treatment delay, loss of treatment effect and treatment waning. We apply this to real world examples showing how these change-point models can adequately model the survival function and produce sensible extrapolations, in contrast to other complex parametric models which may fail to do so.

In the second part of the thesis, we describe a general solution for incorporating expert opinion within statistical models. The focus is on expert opinion on the observable space (i.e. quantities such as survival probabilities) rather than the parameter space (i.e. regression coefficients). We introduce the general framework on how to incorporate such opinions into statistical models, providing examples for important classes of statistical models. We then consider a worked example specific to the extrapolation of survival outcomes.

The work carried out in this thesis has been published in the following articles, R-packages and conferences listed below:

1. Cooney P, White A. Direct Incorporation of Expert Opinion into Parametric Survival Models to Inform Survival Extrapolation. Medical Decision Making. 2023;43(3):325-336.

2. Cooney P, White A. Extending beyond Bagust and Beale: Fully Parametric Piecewise Exponential Models for Extrapolation of Survival Outcomes in Health Technology Assessment. Value in Health, In Press: available online.

3. Cooney P, White A. Incorporating Expert Opinion on Observable Quantities into Statistical Models - A General Framework. Under Review in Bayesian Analysis. Submitted May 2023.

4. Cooney P, White A (2023). expertsurv: Incorporate Expert Opinion with Parametric Survival Models. R package version 1.1.0. Available on Comprehensive R Archive Network (CRAN) [link].

5. Cooney P, White A (2023). PiecewiseChangepoint: Bayesian Change-point Analysis. Available on GitHub.[link]

6. Cooney P (2023), Expertsurv: A Shiny Application for Direct Incorporation of Expert Opinion into Survival Models. R in HTA conference. [link]

7. Cooney P (2020), Incorporating clinical opinion into survival extrapolations with visualisations through RShiny. R in HTA conference. [link]

The contributions of this thesis have relevance for a variety of research areas. For HTA of treatments with survival outcomes the thesis uses change-point models to investigate hypotheses, primarily related to constant hazard extrapolation but also more complex phenomenon when jointly modelling the survival of treatment and comparator arms. The work including expert opinion with survival models helps integrate both data and subjective opinions in a principled manner. More generally, the contributions of these thesis are of relevance for other research areas such as reliability analysis and even logistical planning purposes of clinical trials (Fang and Su, 2011). In terms of (particularly Bayesian) statistics, the work in this thesis provides a general framework for incorporating expert opinion with statistical models. This approach while straightforward to implement has been demonstrated on a number of important classes of statistical models. This thesis also presents a novel approach to estimating piecewise exponential models, specifically allowing the number of change-points to be treated as a parameter to be estimated. We also estimate more complex change-point survival models using modern Bayesian software programs. Using the code we have written for these models allows future users to easily extend the methods and avoid focusing on computational concerns relating to estimation of the parameters. The analysis conducted during this thesis can be replicated using R code made freely available at `Github`. Furthermore the fully functioning R-packages (`PiecewiseChangepoint` and `expertsurv`) allow for application of the methods to real-world problems. This include a web application which allows users with very limited programming abilities to conduct elicitation of expert beliefs, incorporate these beliefs with observed trial data and subsequently produce reproducible reports.

1.4 Outline of Thesis

The work carried out in this thesis is structured in eight chapters across three parts. Part I contains the current chapter, while Chapter 2 introduces the relevant concepts of health technology assessment and survival analysis, summarising approaches to extrapolation of survival outcomes. A discussion about the challenges of predicting long-term survival and limitations of existing methods is provided. The subsequent five chapters contain the contributions of this thesis (Chapter 3 to Chapter 8). They are classified in three parts.

Part II includes Chapters 3 to 5. Here, we describe Bayesian methods to estimate change-point survival models.

In Chapter 3 we consider with the constant hazard change-point model reviewing previous literature and motivating the advantages of Bayesian methods for this class of problems. We develop two Bayesian approaches to estimate both the number and change-point locations.

In Chapter 4 we consider the application of the piecewise change-point model developed in Chapter 3 to estimate the time at which a constant hazard appears plausible. We compare those estimates to those obtained when using an alternative methodology described by Bagust and Beale (2014).

In Chapter 5 we describe more complex change-point models which can be estimated using modern Bayesian programs. We apply these change-point models to survival data exhibiting characteristics which alternative survival models fail to model correctly, specifically changes in the relative treatment effects.

Part III contains Chapters 6 & 7 and describes the research on including expert opinion with statistical models. Chapter 6 describes the general framework applied to a variety of statistical problems and discusses considerations around estimating the relative strength of expert opinion compared with observed data.

Chapter 7 describes the inclusion of expert opinion with a variety of survival models, over a number of timepoints, considering how to resolve/combine the opinion of multiple experts, while also assessing the informativeness of an expert's opinion.

Part IV contains one chapter (Chapter 8) and presents the R-packages used to implement the methods described in Parts II and III, primarily those described in Chapters 3 and 7.

Part V contains the conclusions of this thesis and Part VI.

2 Overview of Survival Analysis relevant to HTA

In this chapter we present an overview of health technology assessment and of survival analysis and its applications to health technology assessment. For those readers who are unfamiliar with HTA, an overview is provided in Section 2.1, explaining why long-term predictions of survival are required.

Subsequently, in Section 2.2 we first define the fundamental concepts of survival analysis. Following this, we present the rationale for and a brief review of the different approaches for extrapolating survival outcomes beyond observed data. In particular we focus on the approaches which are improved upon in this thesis.

2.1 Overview of Health Technology Assessment

The primary objective of this thesis is to investigate statistical techniques to inform extrapolation of time-to-event data from clinical trials for the purposes of economic evaluation of therapies. Therefore, it is useful to provide a brief overview of economic evaluation of medical interventions.

In many countries, health is primarily or substantially funded by the government. The basic idea of economic evaluation is to make a value judgement on a project (e.g. making a new pharmaceutical therapy publicly available) involving public expenditure, given a finite budget and other potential investment choices. Other potential investment choices could be other medicinal therapies or more broadly additional health care staff or infrastructure.

Economic evaluation can be defined as a comparison of alternative options in terms of their costs and consequences (Drummond et al., 2015). Costs can be thought of as the value of the resources involved in providing a treatment or intervention; this would invariably include health care resources, and might be extended to include social care resources, those provided by other agencies, and possibly the time and other costs incurred by patients and their families or other informal carers. Consequences can be thought of as the health effects of the intervention. Two of the primary methods of economic evaluation are discussed in the subsequent sections.

2.1.1 Cost-effectiveness analysis (CEA)

In cost-effectiveness analysis we first calculate the costs and effects of an intervention and one or more alternatives, then calculate the differences in cost and differences in effect, and finally present these differences in the form of a ratio, i.e. the cost per unit of health outcome or effect (Weinstein and Stason, 1977). Because the focus is on differences between two (or more) options or treatments, analysts typically refer to incremental costs, incremental effects, and the incremental cost-effectiveness ratio (ICER). Thus, if we have two options a and b, we calculate their respective costs and effects, then calculate the difference in costs and difference in effects, and then calculate the ICER as the difference in costs divided by the difference in effects:

$$ICER = \frac{Cost_a - Cost_b}{Effect_a - Effect_b} = \frac{\Delta Cost}{\Delta Effect}$$

The effects of each intervention can be calculated using many different types of measurement unit. Two diagnostic tests could be compared in terms of the cost per case detected, two blood pressure interventions by the cost per 1 mmHg reduction in systolic blood pressure, and two vaccination options by the cost per case prevented. However, decision-makers will typically be interested in resource allocation decisions across different areas of health care: for example, whether to spend more on a new vaccination programme or on a new blood pressure treatment. Consequently a measure of outcome that can be used across different areas is particularly useful, and the measure that has so far gained widest use is the quality-adjusted life-year (QALY).

2.1.2 Cost-benefit analysis (CBA)

As stated in the previous paragraph, a key distinction between CEA and CBA is the QALY. The QALY attempts to capture in one metric the two most important features of a health intervention: its effect on survival measured in terms of life-years, and its effect on quality of life.

The other key distinction is the concept of monetary value of health. CEA places no monetary value on the health outcomes it is comparing. It does not measure or attempt to measure the underlying worth or value to society of gaining additional health benefits, but simply indicates which options will permit more health benefits to be gained than others with the same resources.

In contrast, CBA attempts to place some monetary valuation on health outcomes as well as on health care resources. If a new surgical procedure reduces operative mortality by 5%, a cost-benefit approach would try to estimate whether each death averted had a value of €5000 or €500,000 or €5 million, and then assess whether the monetary value of the benefits was greater or less than the costs of obtaining these benefits.

Typically in Ireland figures of €20,000 and €45,000 per QALY are defined as the willingess to pay threshold, i.e. how much the decision-maker is willing to spend to get an extra QALY. Although budget constraints are an important consideration, Irish decision makers will typically fund new interventions that fall below this €45,000 per QALY threshold (Health Information and Quality Authority, 2020).

CBA allows for allocative efficiency, and (in theory) prioritize the reimbursement of different therapies across disease areas in terms of their cost-benefit ratio to make sure that the available resource are directed towards the therapies offering the largest health improvements.

2.1.3 Cost-effectiveness plane

When making assessments about CBA (which in a slight abuse of notation is often referred to as cost-effectiveness), a new therapy will be compared against the current standard of care. This can be represented graphically in the form of the cost-effectiveness plane shown in Figure 2.1.

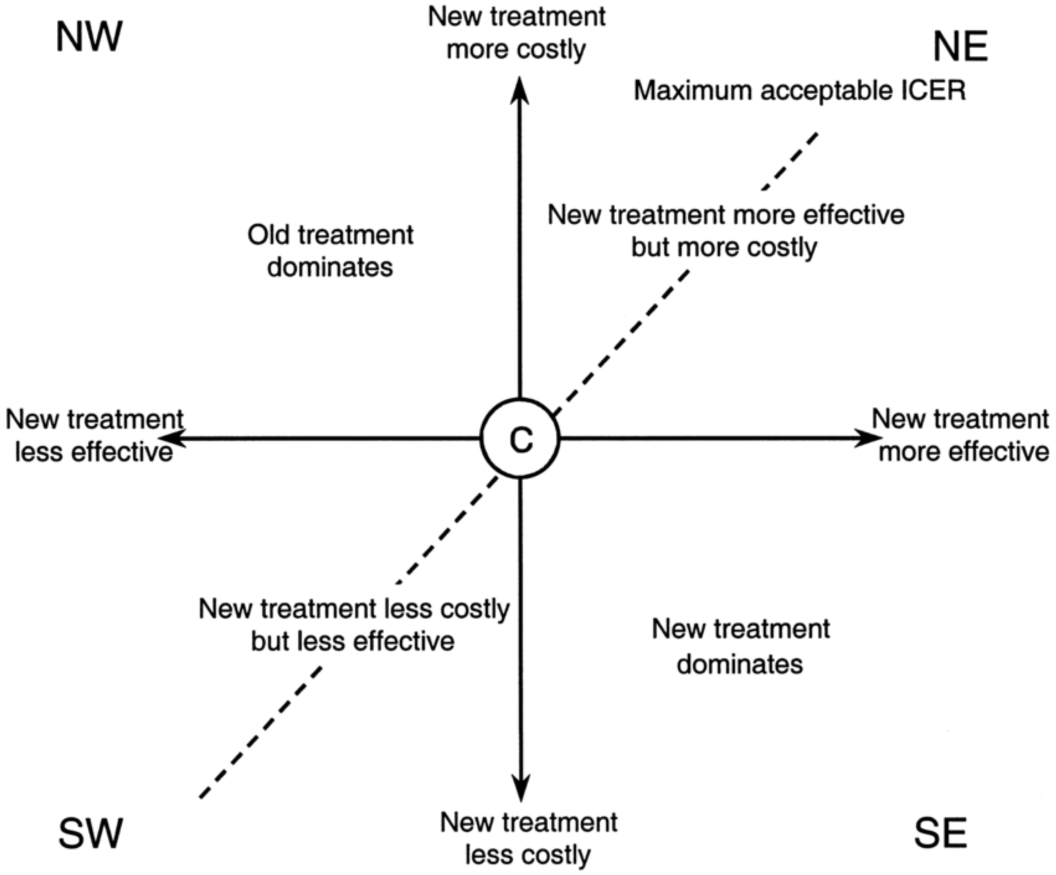


Figure 2.1: Cost-effectiveness plane illustrating the quadrants and their interpretations. Reproduced from Briggs and Tambour (1998).

The most common situations arise in the north-east and south-west quadrants, where the new intervention is more effective but also more costly (the north-east quadrant, quadrant 1), or is less effective but also less costly (the south-west quadrant, or quadrant 3). In these areas of the figure, there is a trade-off between effect and cost: additional health benefit can be obtained but at higher cost (north-east), or savings can be made but only by surrendering some health benefit (south-west). In general if the ICER result is below the cost-effectiveness threshold it will be deemed cost-effective.

2.1.4 Valuing Health - QALY

QALYs are a measure of outcome which typically assigns to each period of time a weight corresponding to the health-related quality of life during that period. Normally the weight 1 corresponds to full health and the weight 0 corresponds to a health state equivalent to dead. Figure 2.2 provides a graphical representation of the QALY approach, in which the life courses of two hypothetical individuals are plotted, with quality of life on the y-axis and time or survival on the x-axis. In this figure, both patients start with similar level of quality of life of 0.8 on a 0–1 scale. Over a period of time the patient not on the intervention has a series of complications which reduces her quality of life with the final one being fatal. In contrast the second patient (on the intervention) experiences complications later than the first patient, including the fatal complication.

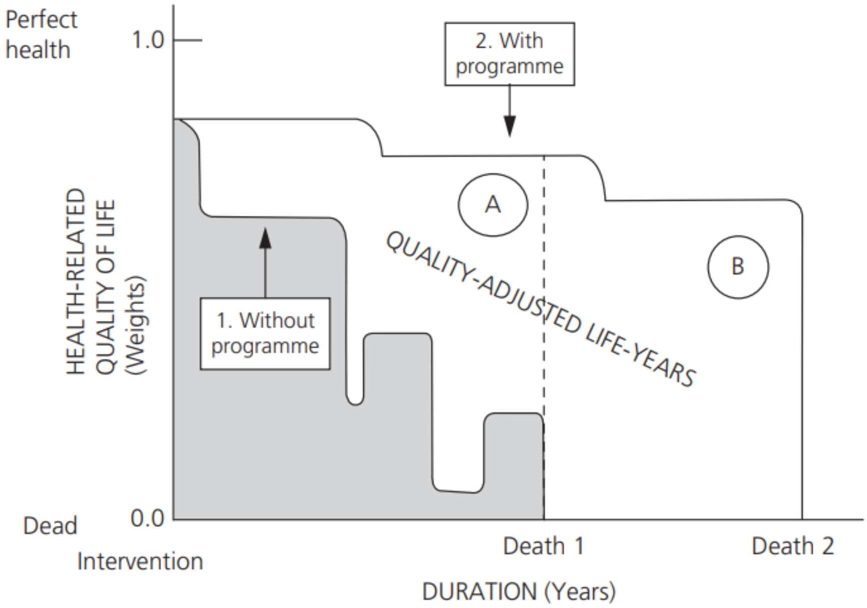


Figure 2.2: Health profile of two individuals in quality and quantity of life dimensions. Reproduced from Drummond et al. (2015).

As discussed by Drummond et al. (2015) it can be seen that it would be possible to measure the difference between these two hypothetical patients in several different ways: by time to first event or complication-free time (a common measure in clinical trials), by time to death, or by number of complications. In this instance any of these would show some benefit to the patient receiving the intervention.

However, all these are partial measures of the differences observed, and measuring the effect of the intervention using any single one of these metrics could be seriously misleading. In contrast, the area under each of the two curves or profiles captures survival as well as the timing and number of non-fatal events and their health impact, and therefore the difference represented by the shaded area is a measure of QALYs gained. Parallels can be made to clinical trials in oncology, where, patients who are progression free have a certain level of quality of life, and which typically is reduced when they progress. Hence it is important to not only be able to estimate expected overall survival, but also expected progression free survival (or more generally any state of health that is expected to be meaningful different in terms of the QALY weight attached to it).

2.1.5 Cost effectiveness models

In order to use the information from a clinical to obtain cost-effectiveness results (i.e. ICER), a decision-analytic model (typically referred to as a cost-effectiveness model) is typically used. It has been noted that relying on a randomized trial as the single vehicle for economic evaluation has a number of limitations (Sculpher et al., 2006). As a result, economic evaluation for decision-making will usually need to draw on evidence from a range of sources. These could include clinical, resource use, and outcome data collected alongside randomized trials, but are also likely to include evidence from other types of studies such as cohort studies and surveys. A decision-analytic model provides a means of bringing together this full range of evidence and directing it at a specific decision problem being addressed by a health system at a given point in time and in a particular jurisdiction.

As discussed by Drummond et al. (2015), cost-effectiveness (CE) models fulfill six main requirements of economic evaluation:

- Comparing all treatment options
- Reflecting all relevant evidence
- Linking intermediate to final endpoints
- Generalizing results to the decision-making context
- Assessment of heterogeneity

and most importantly in the context of this PhD:

- Extrapolating over the appropriate time horizon of the evaluation

In summary, the cost-effectiveness model is the vehicle by which clinical trial data is combined with other information external to the trial, to obtain a cost-effectiveness result which is relevant to the jurisdiction of interest.

2.1.6 Expected Health Benefits and Costs

The expected values of the outcomes from a decision model represents the best estimate of the endpoints of interest for decision-making (Drummond et al., 2015). It is the mean cost and effect, when multiplied by the number of patients treated, gives the total cost and overall health gain for that patient group and therefore, the ICER is based on the Expected (or mean) Cost and QALYs.

This provides the primary motivation for this PhD research. As discussed in Section 2.2 and presented in Equation 2.2, calculation of the mean survival (to which we ascribe QALY weights) requires us to define the survival function across the time horizon of interest. Because of censoring the full survival distribution is typically not available when making an assessment of the cost-effectiveness of a therapy. Therefore, some type of extrapolation of the survival function is required in order to obtain the expected benefits of a treatment. Additionally, because time to treatment discontinuation may also be censored, extrapolation of this function is also required to obtain the expected costs of the intervention.

2.2 Fundamentals of Survival Analysis

This section can be omitted by those familiar with survival analysis.

2.2.1 Study time and patient time

Time-to-event data has two features which require specific statistical methods and the first one is that survival data are generally not symmetrically distributed. Typically, a histogram constructed from the survival times of a group of similar individuals will tend to be positively skewed, that is, the histogram will have a longer tail to the right of the interval that contains the largest number of observations. Secondly the time-to-event data are typically not observed for all observations under study, and the time-to-events for these observations are *censored*. Often this occurs because the data from a study are to be analysed at a point in time when some individuals are still alive. Alternatively, the survival status of an individual at the time of the analysis might not be known because that individual has been lost to *follow-up*.

In a typical study, patients are not all recruited at exactly the same time but accrue over a period of months or even years. After recruitment, patients are followed up until they die, or until a point in calendar time that marks the end of the study, when the data are analysed. Although the actual survival times will be observed for a number of patients, after recruitment some patients may be lost to follow-up, while others will still be alive at the end of the study.

The calendar time period in which an individual is in the study is known as the study time. The study time for eight individuals in a clinical trial is illustrated diagrammatically in Figure 2.3, in which the time of entry to the study is represented by a (•).

This figure shows that individuals 1, 4, 5 and 8 die (D) during the course of the study, individuals 2 and 7 are lost to follow-up (L), and individuals 3 and 6 are still alive (A) at the end of the observation period. As far as each patient is concerned, the trial begins at some time t_0. The corresponding survival times for the eight individuals depicted in Figure 2.3 are shown in order in Figure 2.4. The period of time that a patient spends in the study, measured from that patient's time origin, is often referred to as patient time. The period of time from the time origin to the death of a patient (D) is then the survival time, and this is recorded for individuals 1, 4, 5 and 8. The survival times of the remaining individuals are right-censored (C).

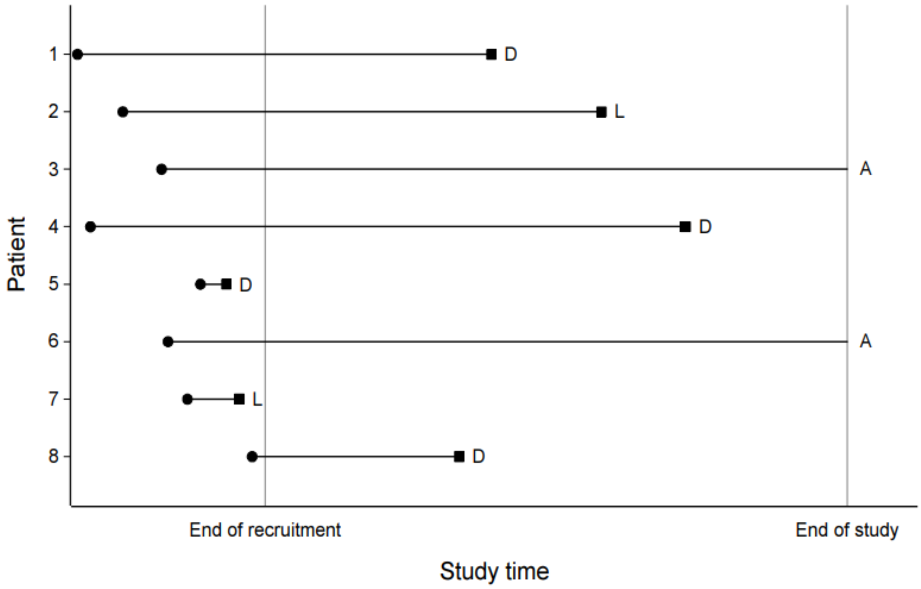


Figure 2.3: Study time for eight patients in a survival study. Reproduced from Collett (2015).

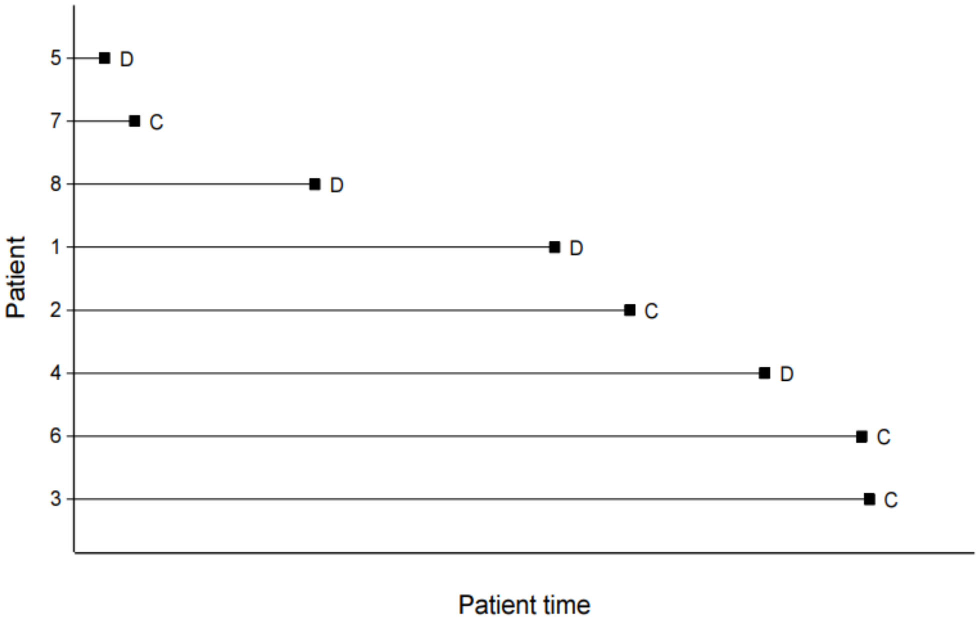


Figure 2.4: Patient time for eight patients in a survival study. Reproduced from Collett (2015).

2.2.2 Survival function

Let T be the random variable for a person's survival time. Since T denotes time, its possible values include all non-negative numbers; that is, T can be any number greater than zero. We let ν denote a $\{0, 1\}$ random variable indicating either failure or censorship. That is, $\nu = 1$ for failure if the event occurs during the study period, or $\nu = 0$

if the survival time is censored. Note that if a person does not fail, that is, does not get the event during the study period, censorship is the only remaining possibility for that person's survival time. That is, $\nu = 0$ if and only if one of the following happens: a person survives until the study ends, a person is lost to follow-up, or a person withdraws during the study period.

We assume that T has a probability density function (p.d.f.) $f(t)$ and cumulative distribution function (c.d.f.) $F(t) = \Pr(T \leq t)$, given the probability that the event has occurred by duration t. The survival function $S(t)$ gives the probability that a person survives longer than some specified time t: that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t (i.e. the complement of the c.d.f function).

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_t^{\infty} f(x)\, dx \tag{2.1}$$

Theoretically, as t ranges from 0 up to ∞, the survival function can be graphed as a smooth curve. As illustrated by the graph, where t identifies the x-axis, all survival functions have the following characteristics (illustrated in Figure 2.5):

1. they are non-increasing; that is, they head downward as t increases;

2. at time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one;

3. at time $t = \infty$, $S(t) = S(\infty) = 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survival function must eventually fall to zero.

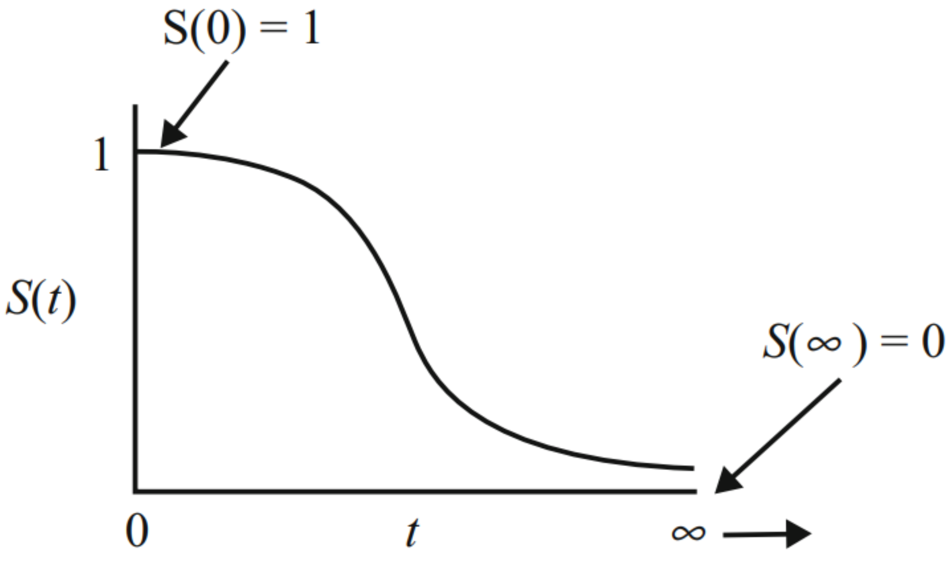


Figure 2.5: Theoretical properties of the survival function. Reproduced from Klienbaum and Klein (2016).

Another useful statistic that can be derived from Survival function is the mean μ or expected value of T. By definition, the expectation of a random variable is calculated by multiplying t by the density $f(t)$ and integrating, so

$$\mu = \int_0^\infty tf(t)dt.$$

Integrating by parts, it can be shown that (for any distribution) that

$$\mu = \int_0^\infty S(t)dt. \tag{2.2}$$

2.2.3 Observed Survival function

When using statistical models to describe survival data we usually consider the probability of survival vs time to be a smooth continuous function (as in Figure 2.5). In practice, when using actual data, we usually obtain graphs that are step functions, as illustrated in Figure 2.6, rather than smooth functions. Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event. The estimated survival function, denoted by a $\hat{S}(t)$ in the graph, thus may not go all the way down to zero at the end of the study.

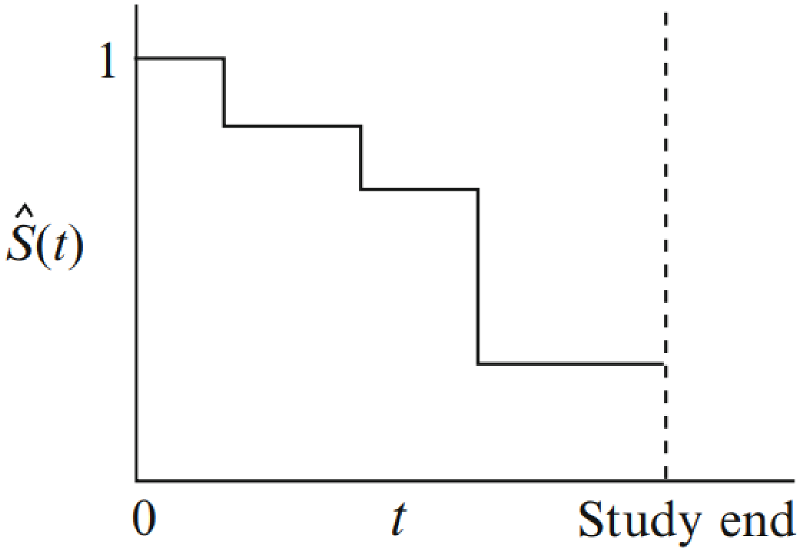


Figure 2.6: Real life survival function. Reproduced from Klienbaum and Klein (2016).

2.2.4 Hazard and Cumulative hazard function

The hazard function, denoted by $h(t)$, is given by the formula: $h(t)$ equals the limit, as t approaches zero, of a probability statement about survival, divided by dt, where dt denotes a small interval of time:

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \tag{2.3}$$

The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $[t, t + dt)$ and $T \geq t$ (which is, of course, the same as the probability that t is in the interval), to the probability of the condition $T \geq t$. The former may be written as $f(t)dt$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.4}$$

From the definition of $S(t)$ in Equation 2.1 it can be seen that $-f(t)$ is the derivative of $S(t)$. Noting that $\int \frac{1}{x} dx = \log |x| + C$ allows us to rewrite the Equation 2.4 as:

$$h(t) = -\frac{d}{dt} \log S(t). \tag{2.5}$$

With some further manipulation of Equation 2.5 the following equation for S(t) is obtained:

$$S(t) = \exp\{-\int_0^t h(x)dx\}. \tag{2.6}$$

Another function that is closely related to the hazard and survival function is the cumulative hazard $H(t)$, which can be considered the sum of the risks you face going from duration 0 to t:

$$H(t) = \int_0^t h(x)dx = -\log S(t). \tag{2.7}$$

An important point to note is that throughout our analysis we assume non-informative censoring. This means that the actual survival time of an individual, t, does not depend on any mechanism that causes that individual's survival time to be censored at time c, where $c < t$. Statistical methods in survival analysis typically make this assumption by default, because analogous to not missing at random (MNAR) longitudinal data, little meaningful analysis can be performed with the introduction of external assumptions/information.

2.3 Parametric Analysis of Right-censored data

In time-to-event analysis we typically deal with right-censored data. Supposing there are n subjects under study, and that associated with the i^{th} individual is a survival time t_i. The t_i's are assumed to be independent and identically distributed (i.i.d) with density $f(t)$ and survival function $S(t)$. The survival time t_i can be censored which we represent by an indicator function ν_i which

$$\nu_i = \begin{cases} 1 & \text{if event} \\ 0 & \text{if censored.} \end{cases} \tag{2.8}$$

Considering a parametric survival model, let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$ be a p-dimensional vector of parameters, the likelihood function the survival model given the observed data D, is

$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} f(t_i)^{\nu_i} S(t_i)^{1-\nu_i} = \prod_{i=1}^{n} h(t_i)^{\nu_i} S(t_i)$.

Because we require predictions for the survival beyond the observed data we require models in which the distribution of the outcome (i.e., the time-to-event) is specified in terms of parameters. Consequently partial likelihood approaches such as the

Cox-proportional hazard model cannot be used as they treat the baseline hazard as a nuisance parameter (See Section 2.4).

In HTA the parametric models which are typically considered are presented in Table D.6. Covariates (typically treatment status) can be included on any of the parameters; however, they are almost exclusively placed on the location parameter. Introducing covariates in this fashion produces parametric models which can be acceleration failure time (AFT) models or proportional hazard models (PH), or neither if placed on the ancillary parameters. The underlying assumption for AFT models is that the effect of covariates are multiplicative (proportional) with respect to survival time, whereas for PH models the underlying assumption is that the effect of covariates is multiplicative with respect to the hazard. For example a Weibull model can be either an AFT or PH model, however, the parameterization is different (see Table D.6).

2.3.1 Frequentist Approach to estimation

In a frequentist approach we seek to maximize the likelihood of the data given the model parameters of dimension m. The parameters which achieve this maximum are denoted as $\hat{\boldsymbol{\theta}} = \text{argmax } L(\boldsymbol{\theta}|D)$. In the presence of a sufficiently large sample size n the standard approximate $1-\alpha$ confidence region for $\boldsymbol{\theta}$ is given by $A = \{\boldsymbol{\theta} : n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^t \hat{\Sigma}_n^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq q_{1-\alpha}\}$, where $q_{1-\alpha}$ is α quantile of the chi-squared distribution with m degrees of freedom and $\hat{\Sigma}_n$ is a consistent estimator for the asymptotic variance which is usually the variance-covariance at $\hat{\boldsymbol{\theta}}$.

A commonly used package implementing a wide range of parametric survival models is the `flexsurv` package (Jackson, 2016). In this package the maximum likelihood of the parameters and associated Hessian matrix (which is then converted to the variance-covariance matrix) are estimated using the `optim` function from the `stats` package (R Core Team, 2021). The default optimization procedure is "BFGS" method, however, other approaches can also be considered (Nocedal and Wright, 2006). Parameter uncertainty for functions of the parameters such as survival and hazard functions are estimated using simulation-based approximations (Mandel, 2013).

2.3.2 Bayesian Approach to estimation

In a Bayesian analysis we assume a prior probability distribution for $\boldsymbol{\theta}$ denoted by $p(\boldsymbol{\theta})$. Knowledge before observing the data as prior information, and to that obtaining after observing the data as posterior information. The posterior distribution is then $\pi(\boldsymbol{\theta}|D) \propto L(\boldsymbol{\theta}|D)p(\boldsymbol{\theta})$. Central to the Bayesian philosophy is that all unknown quantities are described probabilistically, even before the data has been observed. In frequentist statistics, parameters cannot be random variables, and it is not legitimate to make probability statements about them. The prior and posterior distributions gives the full probability distribution of the parameters before and after observing the data, and statements such as "the probability that the parameter is within the interval (x, y) is $z\%$" are valid in contrast to the frequentist method. It is the treatment of unknown parameters as random variables or fixed values that is defining characteristic between Bayesian and frequentist inference, rather than the prior distribution which was typically the point of criticism of Bayesian methods (O'Hagan, 2008).

Much of the controversy about regarding to Bayesian inference has centered on the subjectivity of the prior distribution and that different priors would result in different posteriors. A counter argument is that in every field of science and research there are differences of opinion over topics of current interest. Through use of priors the Bayesian approach allows us to formally include these opinions. These opinions can be on the parameters of a parametric model or on potentially observable quantities that arise from a model. For example, rather than eliciting a distribution for a binomial probability parameter directly "parameter space", an expert may instead be asked to consider hypothetical observations, which are then used to infer a subjective probability distribution for the probability parameter "observable space". It should be noted that although expert opinion is more commonly associated with the Bayesian paradigm it is also possible to include certain beliefs under a frequentist approach primarily through the use of psuedo-observations.

Bayesian approaches are more computationally intensive than frequentist approaches, however, there are a number of different modern software programs available to reliably estimate the posterior distribution. One available R-package with a focus on Bayesian survival analysis is `survHE` (Baio, 2020) which can fit many of the survival models considered in the `flexsurv` package.

For the models estimated using the Bayesian approaches we consider a variety of different approaches. In Chapter 3 we use Gibbs sampling to find the parameters of a piecewise exponential model assuming a fixed number of change-points. We then generalize this model to include a Metropolis-Hasting step when moving between models with different change-point numbers. In Chapter 8 we describe how we optimized estimation of these models using bespoke code written in the C++ programming language.

In later chapters we estimate Bayesian models using statistical programs such as JAGS and Stan (Plummer, 2003; Stan Development Team, 2020). This allows us to construct the model with purpose built robust, validated software rather than writing a bespoke sampling scheme. Models in JAGS are primarily estimated using slice sampling (Neal, 2003) (except when conjugate distributions are available) and models in Stan use Hamiltonian Monte-Carlo (HMC).

Computational methods for Bayesian analysis is a deep topic of research and outside the scope of this thesis. Accessible explanations of Random-Walk Metropolis-Hasting, Gibbs sampling, slice sampling and HMC are provided by Bishop (2006). These methods and other modern advancements of these techniques are documented in detail by Brooks et al. (2011).

Typically, there is a trade-off between the computational complexity of the method and efficiency (per step/simulation) at which the sampler explores the posterior distribution, however, this is situation specific. Generally speaking, Random-Walk Metropolis is the most general approach, however, the choice of step size for the proposed parameters has substantial impact on the acceptance rate and correlation of the accepted parameters. Gibbs sampling is a special case of the Metropolis-Hasting algorithm in which every proposed move is accepted as sampling is done from a conjugate distribution.

Slice sampling is related to Gibbs sampling (as each move is accepted) which does not require conjugacy but requires a number of calculations to define the “slice” from which the parameters are sampled from.

HMC uses information about the gradient of the log probability distribution as well as about the distribution itself to produce more efficient sampling than that typically achieved by Random-Walk Metropolis-Hastings and scales well to high dimensional problems.

Bayesian Software:

In this thesis, two different Bayesian computational software programs are used. The primary software used is Stan, which uses Hamiltonian Monte Carlo (HMC) for computing the posterior distribution. Generally speaking, Stan has been used to estimate a wide variety of statistical models, including models with relatively high numbers of parameters. Estimating the posterior using HMC does come at the cost of increased computation time per sample from the posterior. While HMC can compute the posterior distributions for a wide range of statistical models, there are certain models that we could not program in Stan. Specifically, change-point models with discontinuities in the likelihood result in HMC failing to explore the posterior.

In these instances, we considered an alternative software known as JAGS. The JAGS program primarily uses slice sampling, but when the appropriate likelihood/prior combination is used, the program will use Gibbs sampling.

Although the estimation techniques are different the workflow for each software program is similar. A model script is defined by the user, in which the user expresses the joint relationship between all known (often data) and unknown quantities (parameters) in a model through a series of simple local relationships. Doing so can be more straightforward that directly specifying the complete likelihood function, although this is also a possibility (at least indirectly). The model script and data are then supplied as arguments to an R function which compiles the model in C++ for computational efficiency. Statistical inference is conducted and samples from the posterior distribution are returned. Although there are differences in model syntax, changing an existing model file so that it can be estimated by another program is often straightforward.

The adaptability of these programs is evident in how they handle the methodology for integrating expert opinion described in this thesis. Although this approach necessitates specifying a loss function, it can be seamlessly applied in both Stan and JAGS with minimal adjustments to the standard model files.

2.3.3 Goodness of Fit statistics for Frequentist and Bayesian Models

There are a variety of criteria which can be used to assess the relative fit of alternative models to the data. Generally, they make a trade off regarding how well the model fits the data along with a penalty term accounting for the number of parameters used to estimate the model.

Under the frequentist approach, commonly used criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Akaike, 1998; Schwarz, 1978). AIC includes the log-likelihood at the MLE ($\log L(\hat{\theta}|D)$) multiplied by -2 with the penalty term of 2 times the number of parameters:

$$\text{AIC} = -2\log L(\hat{\boldsymbol{\theta}}|D) + 2p.$$

BIC is similarity defined expect for the penalty term for which n is the number of observations:

$$\text{BIC} = -2\log L(\hat{\boldsymbol{\theta}}|D) + \log(n)p.$$

The interpretation of n is unclear as censored events contribute "less" to the likelihood i.e. only the survival function and not the hazard function. Although Volinsky and Raftery (2000) show that the number of uncensored events is a more appropriate choice (at least for the exponential model), using the BIC function from the `stats` package (R Core Team, 2021) will include censored events in the calculation of n.

For Bayesian analysis one common criterion is Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). DIC uses the value of the log-likelihood at the posterior mean of the parameters $\log L(\bar{\boldsymbol{\theta}}|D)$, adjusting for the effective number of parameters p_D:

$$\text{DIC} = -2\log L(\bar{\boldsymbol{\theta}}|D) + 2p_D.$$

The number of parameters p_D is defined as two times the log-likelihood at the posterior mean of the parameters minus the average log-likelihood over the posterior distribution. Therefore, p_D is calculated using simulations $\boldsymbol{\theta}^{\{s\}}$, $s = 1, \dots, S$ as:
$p_D = 2(\log L(\bar{\boldsymbol{\theta}}|D) - \frac{1}{S}\sum_{s=1}^{S} \log L(\boldsymbol{\theta}^{\{s\}}|D))$. DIC can be seen as a generalization of Akaike's criterion: for models with weak prior information, $\bar{\theta} \approx \hat{\theta}$, and hence $p_D \approx p$ and DIC $\approx$ AIC. DIC has two primary limitations, namely, that it is not invariant to parameterization of the model and can estimate a negative value for the effective number of parameters. A negative value for the effective number of parameters occurs when the posterior mode of the parameters is far from the posterior mean (which can occur in change-point models that we investigate in this thesis).[1]

For these reasons in this thesis we prefer Widely Applicable Information Criterion (WAIC) when assessing goodness of fit in a Bayesian context (Watanabe, 2010). WAIC is constructed as:

$$\text{WAIC} = -2llpd + 2p_{\text{WAIC}}.$$

The lppd (log pointwise predictive density) is defined as the expected value of the likelihood for each individual were D_i is the data for individual i (across the posterior

[1] There is an alternative definition of the effective number of parameters which ensures the number of parameters is positive.

simulations), the logarithm is then taken and summed:

$$llpd = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} L(\boldsymbol{\theta}^{\{s\}}|D_i)\right).$$

The number of parameters p_{WAIC} is defined as follows:

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{S}\sum_{s=1}^{S} L(\boldsymbol{\theta}^{\{s\}}|D_i)\right) - \frac{1}{S}\sum_{s=1}^{S}\log L(\boldsymbol{\theta}^{\{s\}}|D_i)\right).$$

Given a matrix of the log-likelihood for each observation for all simulations the `waic` function from the `loo` package calculates the WAIC (Vehtari et al., 2020).

Another criterion, Pseudo-Marginal Likelihood (PML) described by A. E. Gelfand (1994) has been used for the purposes of averaging competing models in HTA (Jackson et al., 2010), and is defined as:

$$PML = \sum_{i=1}^{n}\log\left(\frac{S}{\sum_{s=1}^{S}\frac{1}{L(\boldsymbol{\theta}^{\{s\}}|D_i)}}\right).$$

In other to place PML on the same scale as the other criterion we typically present $-2 \times$ PML. For $-2 \times$ PML and all other criteria detailed above a lower value indicates better fit.

Note that DIC, WAIC, and $-2 \times$ PML all are motivated as approximations to the out-of-sample predictive accuracy of the model. In our experience, WAIC and $-2 \times$ PML provide very similar estimates of model fit (e.g., equal to one decimal place as in Table 3.7) for standard parametric survival models. However, this may not be the case for other statistical models.

Another possibility is to evaluate the probability of the data under the model, which is known as the marginal likelihood:

$$\pi(D) = \int_{\theta}\pi(D|\theta)\pi(\theta)d\theta.$$

BIC described earlier is an asymptotic approximation of the marginal likelihood under unit information priors (Schwarz, 1978). In contrast to the other measures, the marginal likelihood does not require an explicit penalty term as the approach compares the average fit of a model. This imposes a "natural" penalty for parameters, because each additional parameter introduces a dimension that must be averaged over. If that dimension introduces substantial parameter space with small likelihood, and little space that improves the likelihood, it will decrease the marginal likelihood.

Because marginal likelihood integrates the likelihood with respect to the prior, the choice of prior will affect the final probability. This can be considered a disadvantage when the choice of prior (in terms of its strength or functional form) is unclear. Kass and Raftery (1995) provide a comprehensive overview of Bayes Factors, which represent the ratios of marginal likelihoods for various models. The authors also offer guidance on specifying priors for models where Bayes Factors are computed, even in cases where prior information about model parameters is scarce.

In Chapter 3 we use Bayes Factors to choose between models with different change-point numbers, also providing an approach which reduces some of the sensitivity of the final result to the choice of the priors on the parameters.

2.4 Cox Semi-Parametric Survival Model

Although parametric survival models are the focus of this thesis (due to their ability to predict long-term survival), semi-parametric Cox models are commonly used to estimate relative treatment effects (i.e. hazard ratios) when are then used in decision modelling. As discussed by Dias et al. (2011) a cost-effectiveness analyses (CEAs) consist of two separate components: a baseline model that represents the absolute natural history under a standard treatment in the comparator set, and a model for relative treatment effects.

The former may be based on trial or cohort evidence, while the latter is generally based on randomised controlled trial (RCT) data. In survival modelling, we may have local registry data which represents the natural history and only wish to use the hazard ratio from the clinical trial. Even if the clinical trial population is different to the natural history population, often the relative treatment effective can remain stable across populations.

The Cox PH model is usually written in terms of the hazard model formula $h(t, \mathbf{Z}) = \mathrm{h}_0(t) \exp\left(\sum_{i=1}^{p} Z_i \beta_i\right)$, with $\mathbf{Z}$ a $1 \times p$ vector which is a collection of explanatory/predictor variables that is being modeled to predict an individual's hazard and β_i the covariate for the ith predictor variable. The Cox model formula states that the hazard at time t is a product of two quantities. The first of these, $\mathrm{h}_0(t)$, is called the baseline hazard function. The second quantity is the exponential expression to the linear sum of $Z_i \beta_i$. An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the explanatory variables. In contrast the exponential expression, involve the explanatory variables but not t (i.e. time-independent). It is possible, nevertheless, to consider explanatory variables which do involve t and are called time-dependent.

Advantages of the Cox model is that the model is a “robust” model, so that the results from using the Cox model will closely approximate the results for the correct parametric model (if such a model exists) (Klienbaum and Klein, 2016). The proportional hazard assumption is also intuitive and extends to multiple variables. Because of these properties this model is extensively used when presenting results from clinical trials and one of the most commonly cited statistical methods (Cox, 1972).

2.5 Methods to Extrapolate Long-term survival outcomes

HTA requires a comparison of the incremental costs and health effects of competing interventions. For oncology treatments with survival endpoints, long-term outcomes are typically uncertain at the time of both regulatory and reimbursement assessment. Extrapolating survival to a lifetime horizon is explicitly recommended by multiple HTA authorities for economic evaluations of oncology drugs to assess the long-term consequences of the compared strategies such as Canadian Agency for Drugs and Technologies in Health (2017); Haute Autorité de Santé (2020); National Institute for Health and Care Excellence (2022) and Pharmaceutical Benefits Advisory Committee (2016).

As mentioned in Chapter 1 the primary challenge in survival analysis in HTA is the selection of an appropriate survival model. Although model averaging across a number of parametric survival models is possible, typically one survival model is chosen as a “basecase” and others considered in scenario-analysis. We describe two commonly cited alternative methodologies to selecting a basecase model to extrapolate time-to-event outcomes along, we also describe further work which expands upon topics mentioned but not fully implemented in the each the methodologies.

2.5.1 Technical Support Documents

The National Institute for Health and Care Excellence (NICE) with makes reimbursement decisions for the English health service is widely considered a world leader in HTA and publishes of detailed guidance on methodology, including Technical Support Document (TSD) series (Stevens and Longson, 2013). TSD 14 by Latimer (2013) provides methodological guidance on extrapolation of survival outcomes and is cited within several HTA guidance documents (Canadian Agency for Drugs and Technologies in Health, 2017; Haute Autorité de Santé, 2020; National Institute for Health and Care Excellence, 2022; Pharmaceutical Benefits Advisory Committee, 2016). Guidance within NICE TSD 14 focuses on considering which parametric models are appropriate given the shape of the hazard and survival functions. The author suggests that the choice can be supported by goodness of fit, both by visually inspecting the log-cumulative hazard plots and by statistical goodness of fit.

Log-cumulative hazard plots presenting an approximately linear trend which may be suggestive of a particular parametric model (i.e. exponential or Weibull), however, it is also worth highlighting that there is considerable subjectivity in the assessment of these plots and there are many different candidate models other than exponential and Weibull models. Measures of statistical fit are often used for model selection as they are quantitative and can compare the full range of parametric models. Because the frequentist approach is the most commonly used in HTAs of survival outcomes AIC and BIC are discussed in most detail. In comparison Bayesian methods are less frequently used in survival analysis for HTA, however, the author did suggest DIC as a possible criterion for model selection. Irrespective of the criterion used, the author emphasize that it will only measure goodness of fit to the *observed* data and may still be a poor fit to the true (unknown) long-term survival. Therefore consideration is also given to more subjective assessments such as clinical plausibility of long-term extrapolations (based on similar treatments in the disease area) and biological hypothesis.

External validity is also emphasized in this TSD by comparing the extrapolated hazards to background mortality and survival with data from other studies, such as longer-term follow-up studies or registries where available or expert opinion. This information includes both "hard" data such as registries or "soft" data such as opinion of the analyst or expert and in some situations both types of evidence can be incorporated in to the analysis. A later NICE TSD 21 Rutherford et al. (2020) focused on a range of advanced survival techniques which were not covered in NICE TSD 14, including spline and piecewise models along with posing a number of research objectives.

In both TSD 14 and 21 there are a number of references to expert opinion, however, details on how this can be incorporated with survival models is lacking and TSD 21 explicitly highlights this topic in its recommendations for research. In TSD 21 piecewise models are discussed at some length and are critiqued as they require justifications for the time-intervals used, however, no guidance is provided on methods to select the appropriate intervals.

2.5.2 Bagust and Beale Framework

An alternative framework to the default use of parametric survival models was suggested by Bagust and Beale (2014), based on their extensive experience in conducting HTA of treatments which potentially extend survival. They suggest examining the clinical trial for biologically plausible hypotheses rather than assuming the data is best described by one of the standard parametric survival models. A more detailed overview of the framework is provided in Appendix B.1, however of particular interest to this thesis is their hypothesis that the trial protocol (or disease processes) may induce transient effects on the hazard of an event. These effects can either suppress or elevate observed hazards e.g., exclusion criteria may ensure that the risk of any patient experiencing a target event (death, disease progression, or acute crisis) is artificially suppressed for several months. Additionally (although not directly described in Bagust and Beale (2014)), this could also occur due to the disease process, such as HIV-1 infected patients treated with anti-retroviral therapy, for whom the risk of an event (death/AIDS) becomes relatively stable after several months of treatment (Egger et al., 2002). After these transient effects dissipate a parsimonious choice is to extrapolate survival using a constant hazard model.

Bagust and Beale suggest visually inspecting the empirical cumulative hazard function, that is, the negative logarithm of the Kaplan-Meier (KM) survival function plotted against time, to identify a timepoint after which there is evidence of a long-term linear trend. Once identified, a constant hazard (exponential) model should be fit to the data after this timepoint, a methodology which we hereafter refer to as Bagust and Beale (B&B) approach. This approach assumes that the hazards observed in the clinical trial (after any transient effects have dissipated) are the best estimate of the predicted long-term hazards. This is in contrast to parametric models (other than the exponential model) which assume that the trend in the hazard function (either increasing or decreasing) is valid beyond the observed trial data.

Several important issues are associated with the B&B approach including:

- identification of the timepoint after which a long-term constant hazard is considered plausible;
- incorporation of uncertainty regarding the location of the timepoint into survival projections;
- objective comparison of the B&B approach with fully parametric models
- assumption of constant hazards may lack face-validy over the course of the lifetime horizon

The subjectivity of the approach is highlighted by Figure 2.7. This shows the cumulative hazard function for data simulated from a piecewise exponential model in which the hazard decreases from 0.75 to 0.25 at time equal 1. Based on visual inspection alone it might be difficult to identify this change-point.

Another related scenario considered by Bagust and Beale (2014) is assuming common hazards for both treatment and control arms after a certain period. Although a potentially parsimonious method of modelling the survival data, the choice of this

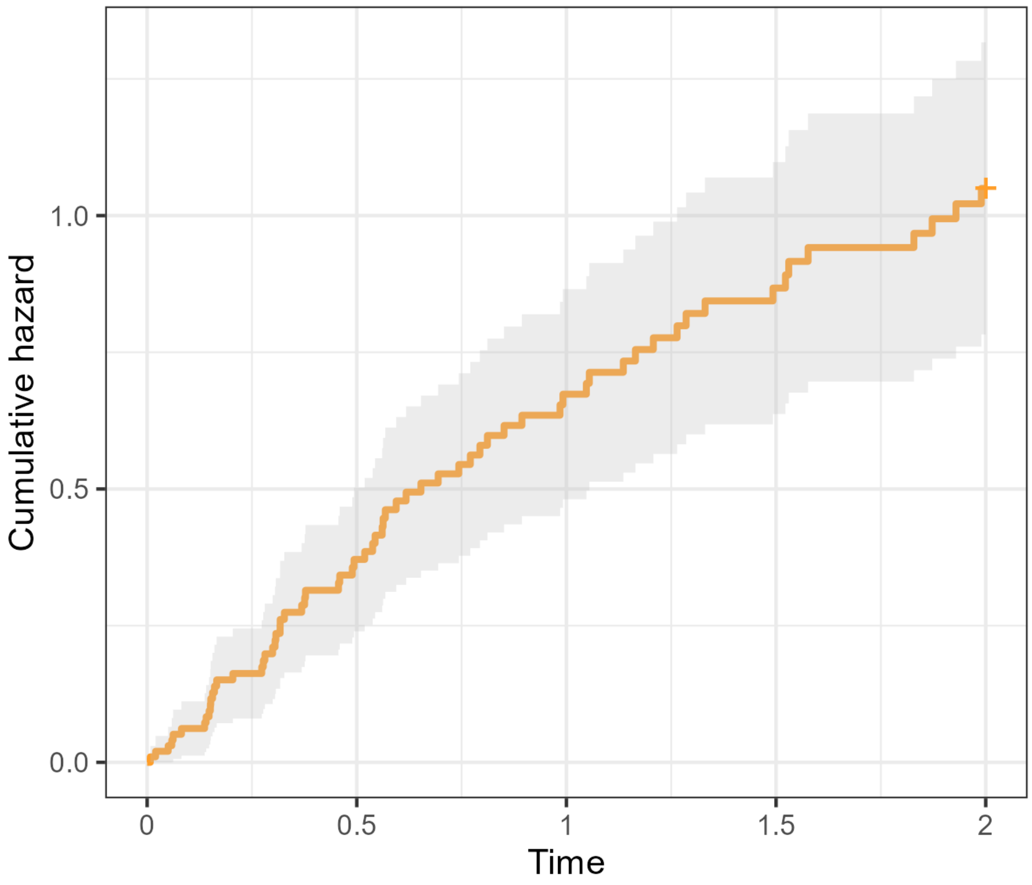


Figure 2.7: Cumulative Hazard plot of data generated from a piecewise exponential model with change-point at time equal to one.

timepoint is again subjective and in Bagust and Beale (2014) is again made through the cumulative hazard function.

2.5.3 Additional approaches and methods to extrapolating survival

As TSD 14 was a guidance document which also sought to direct future research, topics such as inclusion of external information and expert opinion were mentioned without reference to worked examples. Subsequent research such as Guyot et al. (2017) considered using a flexible parametric model fitted to trial data along with registry data and the opinion that both treatments would have equal hazards after a certain period of time. Jackson et al. (2017) describe a variety of different approaches when modelling a treatment and control arm, from the standard proportional hazards, to alternatives such as converging hazards, proportional cause-specific hazards, however, worked examples of these scenarios are not provided. Che et al. (2023) consider another approach where the survival functions (rather than the hazard function) are assumed to converge after a fixed timepoint. One potential criticism of the above approaches is that results are potentially sensitive to the user defined input and do not allow for the observed data to influence these user defined inputs. For example in Guyot et al. (2017), the point at which a common hazard is assumed is a user defined input. In some situations it might be valid to see if the currently observed data supports that assumption, and if so the timepoint at which this is plausible (given the observed data). In both parts of this thesis we will show how the use of change-point models and our approach to incorporating expert opinion into statistical models can allow for the synthesis of the decision maker/expert's belief with the observed data, rather than relying exclusively on one or the other.

2.6 Data used in the Thesis

There are a number of datasets used in the thesis and are sourced from the following:

- Data presented in publications or textbooks

- Pseudo-data generated from published Kaplan-Meier survival functions using the Guyot et al. (2012) algorithm and `digitise` function from the `survHE` package (Baio, 2020).

- Data available in R-packages

These datasets are presented in Table 2.1.

Table 2.1: Overview of datasets used in thesis

Data Description	Chapter Used	Data-source	Reference
Leukaemia Remission Times	Chapter 3	Publication	Matthews and Farewell (1982)
Glioblastoma Survival	Chapter 3	R package	Kosinski and Biecek (2020)
Stanford Heart Survival	Chapter 3	Publication	Miller and Halpern (1982)
Survival Data used Technology Appraisals	Chapter 4	Digitised from publicly available sources	Multiple - See Chapters 4 and 5
E1690 and E1684 trial data	Chapter 5	Sourced from textbook	Ibrahim et al. (2001)
Trial on Exercise programs	Chapter 6	Publication	Littell (1990)
Tisagenlecleucel Survival data	Chapter 7	Digitised from publication	Cope et al. (2019)

2.7 Summary

The challenge of extrapolating survival outcomes beyond observed data is one which is multi-faceted in terms of challenges and potential solutions. Over the course of this thesis we will primarily address the gaps in methodology relating to two topics, improved estimation of the piecewise models (used in the Bagust and Beale approach and more generally to model scenarios relating to treatment effects) and incorporation of expert opinion with survival models.

We propose a statistical method which addresses the limitations of the B&B approach by way of change-point survival models. The approach considered by Bagust and Beale (2014) is non-parametric and emphasizes the use of less complex models for projective modelling of survival outcomes. This is in contrast the approach recommended by Latimer (2013) (TSD 14) who focuses on considering parametric models allow for a wide range of scenarios to be considered and are the default practice in survival modelling in HTA. By considering change-point models which are fully parametric, we will allow comparison with the parametric models considered NICE TSD 14. These models objectively identify the timepoint after which a constant hazard appears plausible allowing for the focus to remain on justification of the underlying model, rather than the value of the timepoint at which constant hazards are assumed plausible.

In Chapter 5 we estimate change-point models with a Weibull likelihood for each segment which can be further extended to other parametric survival models. A key motivation is the observation that in many trials the survival data cannot be adequately described by (even complex) parametric survival models. Analysts typically resort to modelling the data similar to the B&B approach but with a parametric model other than the exponential for extrapolating the long-term survival. Similar to the limitations of the B&B approach described above and as noted in TSD 21, the uncertainty of the point after which the parametric survival model is fit to the data to inform the extrapolated survival is not captured. Therefore the models estimated in Chapter 5 address many of the limitations of piecewise methods described in TSD 21.

The authors of TSD 21 acknowledge that the "approaches that we have outlined have largely tried to capture and extrapolate the marginal hazard and survival functions without trying to compartmentalise the mechanisms driving changes in these functions." A fully parametric change-point model which can capture changes in the model parameters (i.e. shape and scale for Weibull) and the hazard ratios for treatments would capture phenomena such as treatment effect delay and treatment waning along with changes in the disease process which would allow various competing assumptions to be tested.

We also consider the estimation of semi-parametric change-point models. This is important in situations where the relative treatment effect is the focus and the proportional hazards assumption does not hold.

Complementary to this work-stream is the requirement for a more general approach to including expert opinion with survival models. There is a requirement for any opinions to be incorporated in a transparent manner which allows for both the data opinion and expert opinion to influence the result in a manner which is proportional to their relative strengths. Furthermore, efforts should be made to quantify the strength of an elicited opinion relative to the existing data so that experts know the expected impact of their beliefs on the final result. This may allow for appropriate calibration of the expert's opinion at the elicitation stage; rather than re-weighting opinions in a post-hoc manner at the analysis stage.

As HTA is conducted by a variety of stakeholders with varying degrees of programming and statistical knowledge it is important to provide open source software for the proposed methods. The two R packages are described in detail highlight two which implement many of the analysis conducted in this thesis along with a custom built web application which can be used to conduct both elicitation of expert's belief on survival outcomes and analyse the resulting expert opinion with the survival data.

Part II

Change-point Survival Models

3 Constant Hazard Change-point Survival Models

3.1 Introduction

There are a variety of applications for statistical models which assess how the parameters underlying a data generating process may change over time. Wyse and Friel (2010) provide an accessible review of change-point analysis applied to a variety of data types. One particular function which is subject to change is the hazard rate in survival analysis. As described in Chapter 1 the hazard rate quantifies the instantaneous failure rate of a subject who has not failed at a given time point. Because the survival probabilities are directly related to the integral of the hazard function, changes in this function over time are of interest in a variety of situations. Matthews and Farewell (1982) suggest a real world application whereby physicians are interested in determining whether the hazard of relapse in leukaemia is constant or time varying. Another motivation is the extrapolation (or prediction) of survival outcomes for data in which a substantial number of events are unobserved, the primary focus of this thesis.

Both frequentist and Bayesian methods exist for change-point analysis of hazard functions. Frequentist methods primarily consider likelihood ratio, score or Wald tests that are based on analytical approximations for the asymptotic null distribution of the respective test. However, the justification for these limiting distributions often requires some technical assumptions and conditions that are difficult to verify in practice and may not hold for small-to-moderate sample sizes. Even in the presence of larger sample sizes Raftery (1986) notes that likelihood ratio tests will favour the more complex model even if the simpler model fits the data adequately.

In contrast to frequentist methods, the Bayesian approach does not require asymptotics, instead using a set of prior beliefs which are updated using information from an observed sample. This updated belief or posterior probability distribution is used as the basis for inference about the unknown parameters. Bayesian approaches may have advantages in terms of selecting the appropriate number of change-points. Unlike frequentist approaches there are no restrictions on making multiple model comparisons. Bayesian approaches can readily characterize the uncertainty associated with the hazards and the location of change-points. Because Bayesian approaches update probability beliefs based on the observed data, an initial or prior probability is required for each of the parameters. Depending on the strength of the prior belief, inferences may change based on the use of different priors, therefore the influence of a prior should be discussed in Bayesian analyses.

In most Bayesian approaches to hazard change-point detection, the focus is on estimating the location of a known number of change-points. In many of the frequentist approaches, the focus is on testing the alternative hypothesis of a one change-point model versus the null hypothesis of no change-point without consideration for the presence of multiple change-points. For many real world problems the number of change-points in a hazard function is considered unknown, therefore, methods which can estimate the number of change-points and the uncertainty around the number of change-points and their locations would be a useful advancement on current methods.

In the following chapter we present two novel Bayesian approaches to determining the location and number of change-points for a hazard function. Gibbs sampling has been used by Achcar and Loibel (1998) for estimating the location of a single change-point and is straightforward to extend to multiple change-points. Although this approach can characterize the uncertainty in the location of the change-points and the hazards within each interval, it does not directly determine the number of change-points which best describe the data. We propose to address this by calculating the marginal likelihood for a number of plausible change-point models (including a no change-point model). Model selection is based on the Bayes Factors of the competing models using the decision thresholds of Jeffreys (1961).

Another approach involves extending Markov Chain Monte Carlo (MCMC) so that the model dimension is treated as a random variable to be estimated as part of the MCMC procedure. Wyse and Friel (2010) discuss how this method can be implemented for a number of different types of change-point problems which we now extend to hazard functions.

In Section 3.2 we highlight the previous literature for hazard change-point models, specifically for right censored data. In Section 3.3 we describe the Exponential and Piecewise Exponential models. In Sections 3.4 and 3.5 the required notation and describe our proposed statistical models. Section 3.6 presents a simulation study to determine the sample sizes and changes in hazards required for the adequate estimation of the change-point location and frequency. Section 3.7 highlights potential applications of the methods using the datasets introduced in Section 3.7.

3.2 Previous Literature

The majority of the published methods for change-point detection in the hazard function consider the location of a single change-point and assume constant hazards between change-points. A review is provided by Anis (2009) with another earlier review provided by Müller and Wang (1994). Below we provide an overview of some of the most relevant publications from Frequentist and Bayesian perspectives.

3.2.1 Frequentist approaches

Matthews and Farewell (1982) were the first to consider inference on the time of an unknown change-point for a piecewise exponential model. They noted that no-change-point (null) hypothesis (i.e. where the change-point time is equal to 0) is on the boundary of the parameter space. They considered the problem of testing this null hypothesis by deriving a likelihood ratio test statistic.

Nguyen et al. (1984) noted that the likelihood function is unbounded when the hazard before the change-point is greater than the hazard after the change-point. They also note that the likelihood is also unbounded as the change-point tends towards the final observed event (if there are no censored observations after this event). They identified a consistent estimator of the change-point by examining the properties of the change-point likelihood represented as a mixture and does not have the issues with unboundness described above.

Several authors described methods on how to avoid the likelihood becoming unbounded (Matthews and Farewell, 1985; Yao, 1986; Worsley, 1988). Regarding the question of multiple change-points, Goodman et al. (2011) used a Wald type statistic with an alpha spending function to preserve Type 1 error when considering multiple change-point models.

All of the parametric approaches above assume a constant hazard in the intervals between change-points, however, Palmeros et al. (2018) considered a Weibull change-point model meaning that the hazard within each interval can monotonically increase or decrease. Covariates could also be specified within this model, although, no statistical test was presented to test test the hypothesis of a change-point versus no change-point. Gierz and Park (2022) allow for the testing of multiple change-points through the use of a (computationally intensive) bootstrap approximate of the distribution of the likelihood ratio test statistic.

As illustrated by Figure 3.1 the psuedo-likelihood surface[1] of a piecewise exponential

[1] Parameters representing the interval hazards have been maximized conditional on the change-point locations.

change-point mode is non-smooth with many local maxima making optimization using the `optim` function from the `stats` package in R unreliable, especially with higher numbers of change-points. This necessitates searching for the global optimum at many different locations, something which would substantially increase the computational burden of the approach.

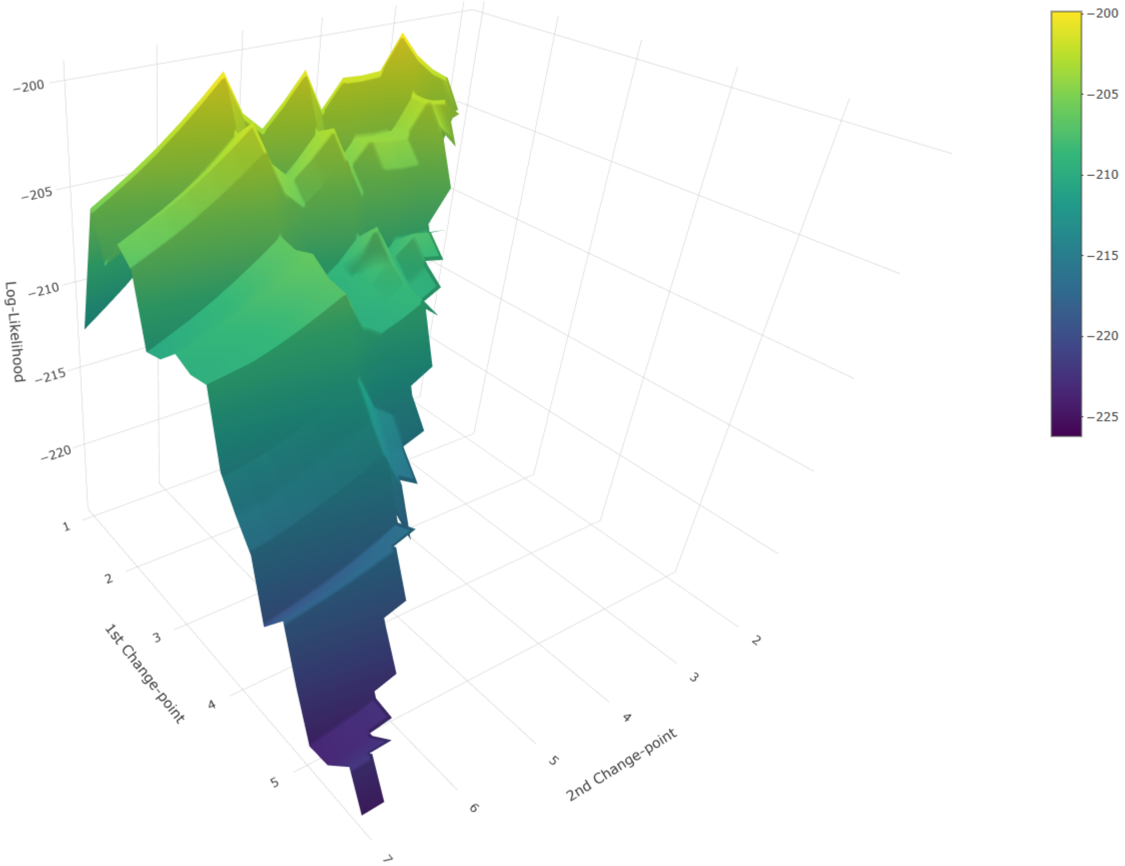


Figure 3.1: Pseudo-Maximized Likelihood surface of a 2 change-point piecewise exponential model

3.2.2 Bayesian approaches

Achcar and Bolfarine (1989) were one of the first to consider hazard change-points from a Bayesian framework. They found the closed form posterior distribution of a change-point, assuming that the change-point occurs at discrete events. Using a non-informative prior for the hazards and a continuous change-point (bounded within a finite interval), Ghosh et al. (1996) discusses an analytic expression of the marginal posterior distribution for a single change-point along with its asymptotic distribution. Karasoy and Kadilar (2007) obtained a Bayes estimate for a change point by considering a prior distribution along with the least squares method proposed by Gijbels and Gurler (2003). Ghosh and Ebrahimi (2008) proposed a Bayes estimator based on a continuous change-point. Achcar and Loibel (1998) presented a Gibbs sampler for a one change-point model, again assuming that the change-point occurs at discrete events. The above Bayesian methods do not consider if the assumption of a change-point is appropriate given the data and estimate the parameters given the presence of a change-point. Yao (1987) proposed a Bayesian test which can test the hypothesis of a change-point versus no change-point, however, it is less efficient than the score test proposed by Matthews et al. (1985).

Kim et al. (2020) use a stochastic approximation Monte Carlo algorithm to identify which particular number and location of change-points gives the highest log-posterior values. Similar to the collapsing model presented in this chapter, they allow the sampler to move between different change-point models as part of the estimation procedure, however, they do not present the relative probabilities of models with different numbers of change-points. Another approach by Chapple et al. (2020) also allows for moves between different change-point models using a technique called Reversible Jump Markov Chain Monte Carlo (RJMCMC), of which our collapsed approach is a special case. In this chapter we also present a Gibbs sampler similar to Achcar and Loibel (1998) to estimate models with multiple change-points and extend this approach to select the best fitting model.

3.3 Exponential and Piecewise Exponential Survival Models

3.3.1 Exponential model

The density function for an exponential distribution is $f(t) = \lambda \exp\{-\lambda t\}$ with support $t \in [0, \infty)$. The support is the range of values that the random variable t may take and because the exponential model (like the other distributions considered in this thesis) only allows positive numbers, it is appropriate for modelling survival times. This is the simplest possible survival distribution as it assumes a constant risk over time, so the hazard is $h(t) = \lambda$ for all t. The corresponding survival function is $S(t) = \exp\{-\lambda t\}$. Taking the product of the hazard function and the survival function produces the density function $f(t) = h(t)S(t)$ which is simply a rearrangement of Equation 2.4 and holds for all survival distributions.

Consider a sample of n observations of survival times $\mathbf{t}_{1:n} = (t_1, \dots, t_n)$ being time ordered, some of which may be censored. The likelihood function may be written as

$$L(\mathbf{t}_{1:n}|\lambda) = \prod_{i=1}^{n} \lambda^{\nu_i} S(t_i),$$

with $\nu_i = 1$ if the subject failed and 0 if censored.

Taking the natural logarithms, and noting that the natural logarithm of the survival function is equal to the negative cumulative hazard function $H(t)$, we obtain the log-likelihood function

$$\log L(\lambda|\mathbf{t}_{1:n}) = \sum_{i=1}^{n} \nu_i \log \lambda - H(t_i).$$

Letting $D = \sum_{i=1}^{n} \nu_i$ denote the total number of observed deaths, and $T_{\text{total}} = \sum_{i=1}^{n} t_i$ denote the total observation (or exposure) time, we can rewrite the log-likelihood as a function of these totals.

Lemma 1. *The log-likelihood of an exponential model is:*
$\log L(\lambda|\mathbf{t}_{1:n}) = D \log \lambda - \lambda T_{total}$.
Because the hazard is constant for all t, the cumulative hazard is the integral of a constant and is λt_i, therefore d_i and t_i can be replaced with their sums.

Exponentiating the log-likelihood provides the likelihood $L(\lambda|\mathbf{t}_{1:n}) = \lambda^D \exp^{-\lambda T_{\text{total}}}$.

This distribution plays a central role in survival analysis, although it is potentially too simple to be useful in applications. Therefore, an extension to the exponential model which allows the hazard to change at various intervals called a piecewise exponential model is discussed in the subsequent paragraph.

3.3.2 Piecewise Exponential Model

A change-point occurs at observation q if $t_1, \ldots, t_q$ are generated differently to $t_{q+1}, \ldots, t_n$. In a piecewise constant model with one change-point, this requires that the segments $\mathbf{t}_{1:q}$ and $\mathbf{t}_{q+1:n}$ have a constant hazard within the segment, but independent hazards between segments. It is assumed that the change-points occur at a particular event time (and not a censoring time). Multiple change-points at specific event times can be denoted as a vector $\boldsymbol{\tau}_{1:k}$, with these k change-points splitting the data into $k+1$ segments. The likelihood of the piecewise exponential model can be formulated as follows

$$L(\boldsymbol{\tau}_{1:k}, \boldsymbol{\lambda}_{1:k+1}|\mathbf{t}_{1:n}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k+1} \lambda_j^{\delta_{ij}\nu_i} \exp\left\{ -\delta_{ij} \left[\lambda_j(t_i - \tau_{j-1}) + \sum_{g=1}^{j-1} \lambda_g(\tau_g - \tau_{g-1}) \right] \right\} \right\} \tag{3.1}$$

with $v_i = 1$ if the ith subject was observed to have an event and 0 otherwise (censored). If the ith subject's time (either censored or an event) is within the jth interval (mathematically $t_i \in (\tau_{j-1}, \tau_j]$), $\delta_{ij} = 1$ and 0 otherwise.

The log-likelihood is then

$$\log L(\boldsymbol{\tau}_{1:k}, \boldsymbol{\lambda}_{1:k+1}|\mathbf{t}_{1:n}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k+1} \delta_{ij} v_i \log(\lambda_j) - \delta_{ij} \left[\lambda_j (t_i - \tau_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (\tau_g - \tau_{g-1}) \right] \right\}, \quad (3.2)$$

with $\sum_{j=1}^{k+1} \delta_{ij} v_i \log(\lambda_j)$ representing the contribution of the log hazard function for an individual (across the segments) and $\sum_{j=1}^{k+1} \delta_{ij} \left[\lambda_j (t_i - \tau_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (\tau_g - \tau_{g-1}) \right]$ the individual cumulative hazard function.

By omitting the potential for covariates and restricting ourselves to discrete change-points, it should be noted that there is no loss of information in recasting the time ordered data as times between individual event times. We let d be the number of event times and $n - d$ right censored survival times. For notational ease, we assume here that only one individual dies at each time, so that there are no ties in the data, however, the model implementation allows for tied events. Denote the ordered distinct survival times by $x_1, x_2, \dots, x_d$, so that x_i is the i^{th} ordered survival time. The set of individuals who are at risk at time x_i will be denoted by $\mathcal{R}_i$ (the risk-set), so that $\mathcal{R}_i$ is the set of individuals who are event-free and uncensored at a time just prior to x_i. $|\mathcal{R}_i|$ is the cardinality or number of individuals in the set. We define y_i as the total (sample) time between events $i - 1$ and i as

$$y_i = (x_i - x_{i-1}) \times |\mathcal{R}_i| + \sum_{j=1}^{n} I(v_j = 0, x_{i-1} < t_j < x_i) \times (x_i - t_i).$$

This is the composed of the difference between event times multiplied by the risk set at the event time plus the difference between any censored observations and the previous event time x_{i-1}, provided they occurred within the interval (x_{i-1}, x_i).

We can re-express the likelihood of the piecewise exponential model in terms of $\mathbf{y}_{1:d}$. Let $\mathbf{s}_{1:k}$ be a vector representing the number of events which have occurred at each of the elements of $\boldsymbol{\tau}_{1:k}$, with $s_0 = 0$ and $s_{k+1} = d$. The likelihood of interval j is $\lambda_j^{s_j - s_{j-1}} \exp\left\{ -\lambda_j \sum_{i=s_{j-1}+1}^{s_j} y_i \right\}$. Censored observations are also allowed, providing exposure time within intervals without an event. The likelihood is then

$$L(\mathbf{s}_{1:k}, \boldsymbol{\lambda}_{1:k+1}|\mathbf{y}_{1:d}) = \prod_{j=1}^{k+1} \left[\lambda_j^{s_j - s_{j-1}} \exp\left\{ -\lambda_j \sum_{i=s_{j-1}+1}^{s_j} y_i \right\} \right].$$

3.4 Estimation using Gibbs Sampler

3.4.1 Gibbs sampler implementation

We propose a Gibbs sampler to find the location of a known number of k change-points with the data recast as a vector of time between events. Noting that a gamma distribution ($\mathcal{G}$) is conjugate prior to an exponential likelihood, we can obtain a marginal distribution of the change-point locations by conditioning on the hazards from the previous step. Choosing a gamma distribution for the hazard parameters is crucial because if we assumed another prior, such as a log-normal distribution, we would no longer have a gamma distribution as a posterior to an exponential likelihood. This choice is not limiting in a practical sense, as the gamma distribution is relatively flexible and its expectation and variance can be adjusted by specifying its parameters α and β.

A Directed Acyclic Graph (DAG) representing the dependency between the change-point locations and hazards is presented in Figure 3.2. For interval j, the difference between s_j and s_{j-1} defines the number of events within that interval (as denoted by a double arrow). The sum of the number of events and the hyperparameter α determines the first parameter of the gamma distribution from which the hazard λ is sampled. The change-point locations s_j and s_{j-1} also define the exposure time within the jth interval such that $\sum_{i=s_{j-1+1}}^{s_j} y_i + \beta$ determines the second parameter of the gamma distribution for the hazard. Repeated parts of the graph are be represented using a "plate", as shown for the range of intervals from 1:(k+1) and also for the time between events within in each interval.

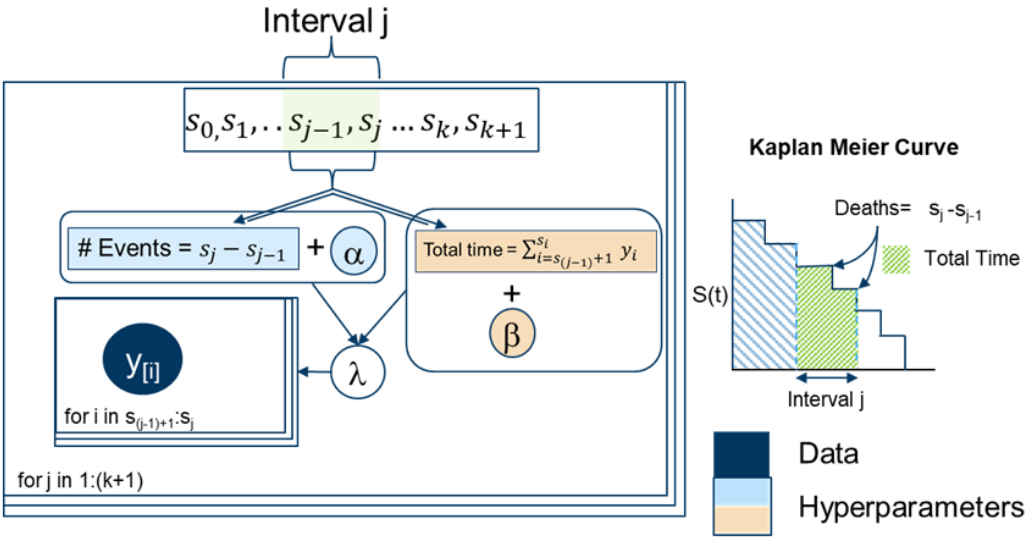

Interval j
$s_0, s_1, .. s_{j-1}, s_j \dots s_k, s_{k+1}$
# Events = $s_j - s_{j-1}$
+
α
Total time = $\sum_{i=s_{(j-1)}+1}^{s_i} y_i$
+
β
λ
$y_{[i]}$
for i in $s_{(j-1)+1}$:s_j
for j in 1:(k+1)
Kaplan Meier Curve
Deaths= $s_j - s_{j-1}$
Total Time
S(t)
Interval j
Data
Hyperparameters

Figure 3.2: Illustration of a Directed Acyclic Graph for Gibbs Change-point Sampler

We define prior probabilities for the change-point locations and hazards as follows:

$$
\begin{aligned}
p(\mathbf{s}_{1:k}|k) &= \binom{n_{\text{events}}-1}{2k+1}^{-1}\prod_{j=0}^{k}(s_{j+1}-s_j-1)\\
\boldsymbol{\lambda}_{1:k+1} &\sim \mathcal{G}(\alpha,\beta).
\end{aligned}
$$

The discrete prior for the change-point locations was presented in Fearnhead (2006) with n_{events} referring to the number of events. This prior has the advantage of ensuring the change-points are not too close together or near the final event (relative to the sample size).

Based on the timescale we define α and β, the hyper-parameters for $\lambda_{1:k+1}$ (a vector of hazards of size $k+1$). If the timescale is years we consider $\alpha = 1$ and $\beta = 1$, while if the timescale is in days we define $\alpha = 1$ and $\beta = 365$. Due to the properties of the gamma distribution, these choices result in a priors with equivalent (scaled) mean and variances.

The unscaled joint posterior density of the hazards and change-points is the product of the likelihood, prior on change-point locations and the prior on the hazards:

$$
\pi(\mathbf{s}_{1:k},\boldsymbol{\lambda}_{1:k+1}|\mathbf{y}_{1:n},k,\alpha,\beta) \propto L(\mathbf{s}_{1:k},\boldsymbol{\lambda}_{1:k+1}|\mathbf{y}_{1:n})p(\mathbf{s}_{1:k}|k)p(\boldsymbol{\lambda}_{1:k+1}|\alpha,\beta).
$$

Conditional on the current change-point locations the hazards $\lambda_{1:k+1}$ are drawn from gamma distributions as follows:

$$
\begin{aligned}
\lambda_1|\alpha,\beta,k &\sim \mathcal{G}(\alpha+s_1,\beta+\sum_{i=1}^{s_1}y_i)\\
\lambda_2|\alpha,\beta,k &\sim \mathcal{G}(\alpha+s_2-s_1,\beta+\sum_{i=s_1+1}^{s_2}y_i)\\
&.\\
&.\\
\lambda_{n+1}|\alpha,\beta,k &\sim \mathcal{G}(\alpha+s_{k+1}-s_k,\beta+\sum_{i=s_k+1}^{s_{k+1}}y_i)
\end{aligned}
$$

The posterior probability of the first change-point is calculated by evaluating the likelihood of all the possible change-points (from 2 to $s_2 - 1$) conditional on the the change-point location s_2 and λ_2. These likelihoods are converted to probabilities by dividing the likelihood of an individual change-point location by the sum of the likelihoods from 2 to $s_2 - 1$ as shown in Equation 3.3 where $j = 1$:

$$f(s_i|\mathbf{y}_{1:n}, s_{j-1}, s_{j+1}, \lambda_j, \lambda_{j+1}, \alpha, \beta, k) = \frac{L(\lambda_j|\mathbf{y}_{s_{j-1}+1:s_i})L(\lambda_{j+1}|\mathbf{y}_{s_i+1:s_{j+1}})p(s_1, ., s_i, .s_k|k)}{\sum_{i=s_{j-1}+1}^{s_{j+1}-1} L(\mathbf{y}_{s_{j-1}+1:s_i}|\lambda_j)L(\mathbf{y}_{s_i+1:s_{j+1}}|\lambda_{j+1})p(s_1, ., s_i, .s_k|k)} \tag{3.3}$$

Based on these probabilities a new location for the first change-point is sampled. Conditional on the newly sampled change-point, the hazards are updated and posterior density of the second change-point is calculated by evaluating the likelihood of the change-points from $s_1 + 1$ to $s_3 - 1$. The second change-point is sampled from this posterior and the process continues until all the change-points have been evaluated.
In summary, the model proceeds as follows:

1. Initialize $\mathbf{s}_{1:k}$ by random draw of size k from 1:(n-1) events.
2. For each iteration, indexed $m = 1, 2, \dots, M$ repeat the following steps:
   (a) For the current values of $\mathbf{s}_{1:k}$, define the number of events and the exposure time within each interval.
   (b) Sample $\lambda \sim \mathcal{G}(\alpha + D, \beta + T)$ for each interval where D is the number of events and T is the total exposure time in the interval.
   (c) For the first change-point, evaluate the likelihood of change-point locations from $s_{(j-1)} + 1 : s_{(j+1)} - 1$, where $j = 1$.
   (d) Sample a new change-point s_j with the probability of each change-point location calculated using Equation (3.3).
   (e) Conditional on the change-points, re-sample the vector of hazards $\lambda_{1:k+1}$ and repeat the previous four steps for the remaining change-points.
3. Increment m.

Model Selection for Gibbs Sampler

The Gibbs sampler provides a posterior distribution of the change-point locations for a given number of change-points, however, we require a method of assessing the appropriate number of change-points. Because the marginal likelihood of the piecewise exponential intervals are available in a closed form (see Appendix A.1), we can calculate the marginal likelihood for the a given change-point model (with parameter set $\boldsymbol{\theta}_k$ for a given values of the hyperparameters γ) denoted as $\pi(\mathbf{y}_{1:n}|\gamma, k) = \int_{\boldsymbol{\theta}_k} L(\mathbf{y}_{1:n}|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\gamma)d\boldsymbol{\theta}_k$. This is achieved by simulating a large number of change-point values from the prior for the change-points. We then calculate the marginal likelihood for each configuration of change-points simulated from the prior and then average the result. This expected marginal likelihood can then be used to compute the Bayes Factor. This can be thought as the magnitude of the evidence for Model 1 over Model 2 where the models differ with respect to their change-point numbers:

$$BF_{1,2} = \frac{\int_{\boldsymbol{\theta}_1} L(\mathbf{y}_{1:n}|\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1|\gamma)d\boldsymbol{\theta}_1}{\int_{\boldsymbol{\theta}_2} L(\mathbf{y}_{1:n}|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2|\gamma)d\boldsymbol{\theta}_2}.$$

The marginal likelihood is the mean likelihood obtained by averaging the likelihood across all parameter values and weighted by the parameter prior. This Bayesian averaging is exactly how Bayes Factor's avoids overfitting, that is, by selecting the model with the highest mean likelihood value, instead of the one with the highest maximum likelihood value.

Bayes Factor are transitive in that multi-way comparisons are relative. So if we have $BF_{1,2}$ and $BF_{2,3}$ then:

$$BF_{1,2}BF_{2,3} = BF_{1,3},$$

which is useful for multiple model comparisons using the same data.

Regarding interpretation of Bayes Factors, Jeffreys (1961, p.432) propose the following criteria in Table 3.1, where Model 2 is the more complex model and Model 1 is H_0.

Table 3.1: Criteria for model comparisons when using Bayes Factors

$\log_{10}(B_{21})$	B_{21}	Evidence against Model 1 (H_0)
0 to 1/2	1 to 3.2	Minimal
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

If we wish to compare across a number of change-point models and select the least complex model (in terms of change-point numbers) with substantial evidence (i.e. $B_{21} \geq 3.2$) we propose the following steps.

1. Identify the model with the highest mean marginal likelihood and discard the more complex models.

2. Compute the logarithm (to the base 10) of the mean marginal likelihood for each model.

3. For each of the simpler models subtract the log mean marginal likelihood from the model with the highest mean marginal likelihood.

4. Check if the difference for any of these models is ≤ 0.5. If yes, choose the simplest model from this set as the optimal model. If not, the chosen model is the one with the highest mean marginal likelihood.

3.5 Estimation using Collapsing Change-point Approach

Markov chain samplers that jump between models with different numbers of change-points allow us to estimate posterior probabilities for candidate models while also estimating the location of change-points within each model. Introducing priors for the change-point numbers, change-point locations and hazards $p(k|\xi)$, $p(\mathbf{s}_{1:k}|k)$, $p(\boldsymbol{\lambda}_{1:k+1}|\alpha,\beta,k)$ respectively, means that we can treat the number of change-points k as a random quantity to be inferred. The model posterior then becomes

$$\pi(k,\mathbf{s}_{1:k},\boldsymbol{\lambda}_{1:k+1}|\mathbf{y}_{1:d},\alpha,\beta,\xi) \propto L(\mathbf{s}_{1:k},\boldsymbol{\lambda}_{1:k+1}|\mathbf{y}_{1:d})p(\mathbf{s}_{1:k}|k)p(\boldsymbol{\lambda}_{1:k+1}|\alpha,\beta,k)p(k|\xi). \tag{3.4}$$

Following the approach outlined by Wyse and Friel (2010), if we regard the hazards $\boldsymbol{\lambda}_{1:k+1}$ as nuisance parameters, the posterior density of the change-point number and their respective locations is proportional to

$$\pi(k,\mathbf{s}_{1:k}|y,\alpha,\beta,\xi) \propto \prod_{j=1}^{k+1} \pi(\mathbf{y}_{s_{j-1}+1:s_j}|\mathbf{s}_{1:k},\alpha,\beta)p(\mathbf{s}_{1:k}|k)p(k|\xi)$$

where $\pi(\mathbf{y}_{s_{j-1}+1:s_j}|\mathbf{s}_{1:k},\alpha,\beta)$ denotes the marginal likelihood of the j^{th} data segment. Adopting a common, independent Gamma prior $\lambda_j \sim \mathcal{G}(\alpha,\beta)$ for $j = 1,\dots,k+1$ makes this quantity straightforward to calculate; see Appendix A.1 for full details.

Because the marginal likelihood of each data segment is available in closed form, a switch from k to $k+1$ change-points, or vice-versa, does not require the design of a bijective function between support subspaces. Therefore, this model is a special case of a RJMCMC. Changes to the change-point number are proposed and accepted with Metropolis-Hastings probability $\min(1, A)$ where

$$A = \frac{\pi(k+1,\mathbf{s}'_{1:k+1}|\mathbf{y}_{1:d},\alpha,\beta,\xi)}{\pi(k,\mathbf{s}_{1:k}|\mathbf{y}_{1:d},\alpha,\beta,\xi)} \times \frac{P(k+1,k)}{P(k,k+1)}. \tag{3.5}$$

The ratio of the marginal likelihoods is straightforward to compute and can be expressed as

$$\frac{\pi(k+1,\mathbf{s}'_{1:k+1}|\mathbf{y}_{1:d},\alpha,\beta,\xi)}{\pi(k,\mathbf{s}_{1:k}|\mathbf{y}_{1:d},\alpha,\beta,\xi)}=\frac{p(k+1|\xi)}{p(k|\xi)}\frac{p(\mathbf{s}'_{1:k+1}|k+1)}{p(\mathbf{s}_{1:k}|k)}\frac{\pi(\mathbf{y}_{s_{j-1}+1:s'_j}|\alpha,\beta)\pi(\mathbf{y}_{s'_j+1:s'_{j+1}}|\alpha,\beta)}{\pi(\mathbf{y}_{s_{j-1}+1:s_j}|\alpha,\beta)}, \qquad (3.6)$$

where the location of the additional change-point is denoted by $s_j^{'}$. When adding a change-point in the proposal step, one of $d-k-1$ points where there could be a change-point are randomly selected. If this point occurs in segment j, segments $\mathbf{y_{s_{j-1}+1:s'_j}}$ and $\mathbf{y_{s'_j+1:s'_{j+1}}}$ are obtained, from which we calculate the marginal likelihoods and prior densities in Equation 3.6. When deleting a change-point, one of the k change-points are randomly selected and $y_{s_{j-1}+1:s_j}$ becomes the new data segment where $s_j = s_{j+1}$ before deletion.

The probability of adding a change-point for a model with k change-points is a_k, and r_{k+1} is the probability of removing a change-point for a model with $k+1$ change-points. Clearly $r_k = 1 - a_k$, with $r_0 = 0$ and $a_K = 0$, for K the largest change-point number under consideration, with $r_k = a_k$ for the other change-point numbers. The proposal one step transition probabilities for the number of change-points are $P(k, k+1) = \frac{a_k}{d-k-1}$ and $P(k+1, k) = \frac{r_{k+1}}{k+1}$.

Following the change-point number proposal step, a single change-point location is also sampled at each iteration. One of the k change-points is randomly selected, and its location sampled with probability

$$\pi(s_j|\mathbf{y}_{s_{j-1}:s_{j+1}}, s_{j-1}, s_{j+1}, \alpha, \beta, k) \propto \pi(\mathbf{y}_{s_{j-1}+1:s_j}|s_{j-1}, s_j, \alpha, \beta)\pi(\mathbf{y}_{s_j+1:s_{j+1}}|s_j, s_{j+1}, \alpha, \beta)p(s_{1:k}|k),$$

for $s_j = s_{j-1}+1, \dots, s_{j+1}-1$.

Regarding priors we assign a Poisson(ξ) for the number of change-points k. In the examples that follow, we set $\xi = 1$. The prior for the change-point locations is as in Section 3.4.1. For the prior for each hazard, $\pi(\lambda_j|\alpha,\beta)$, we set $\alpha = 1$, and the expected value for $\beta = 1$ in the case that the timescale was in years, and $\beta = 365$ or 12 for timescales in days or months respectively. We discuss hyperpriors for β in which provides more robust inferences in Appendix A.1.1. Although we integrate out the hazard parameters λ from the model during this estimation scheme, it is possible to estimate the hazards for a given change-point model using the already sampled change-point locations by simulating draws from the conditional distribution $\pi(\lambda_j|\mathbf{y}_{1:d}, s_j, s_{j-1}, \alpha, \beta)$, for $j = 1, \dots, k+1$. In effect, this introduces an extra sampling step, in which the hazards $\boldsymbol{\lambda}_{1:k+1}$ are "uncollapsed" and sampled at each iteration, before once again being collapsed before the change-point number and locations are sampled, albeit this is done in a post-hoc fashion. The conditional distribution $\pi(\lambda_j|\mathbf{y}_{1:d}, s_j, s_{j-1}, \alpha, \beta)$, has a gamma distribution $\mathcal{G}(\alpha'_j, \beta'_j)$, with shape $\alpha'_j = s_j - s_{j+1} + \alpha$ and rate $\beta'_j = \sum_{i=s_{j_1}+1}^{s_j} y_i + \beta$.

3.6 Simulation Study

The technique presented by Castelloe and Zimmerman (2002) was used to determine the appropriate chain length for all scenarios in the simulation study detailed below. This technique is a version of the Potential Scale Reduction Factor (PSRF) modified for RJMCMC samplers. Across many of the example datasets we considered, the PSRF of the change-point number remained below 1.02 after around 100 iterations of the model. Although this measure is not definitive (and is dynamic with respect to chain length), it potentially indicates adequate mixing with respect to the change-point number. Moreover, because Gibbs sampling is employed for the change-point locations and interval hazards, every proposed move is accepted, which typically promotes effective exploration of the posterior.

We ran the model for 20,750 iterations with the first 750 discarded across two chains for each of the simulation studies detailed below. Because the collapsed model is much more (computationally) efficient at calculating the relative probabilities of competing models than estimating all potential models by Gibbs sampling and then calculating marginal likelihoods, the scenarios are estimated using the collapsing model (with both approaches expected to give very similar results).

3.6.1 Assessment of Power and Parameter Estimation

We conducted a simulation study to investigate the accuracy with which the collapsing change-point model estimated the model hazards, identified the locations of the change-points and selected the correct number of change-points. We simulated data from models with $k = 0, 1, 2$ change-points. For each model we varied the sample size and the characteristics of the hazard function. Data from each scenario was simulated 500 times.

We calculated number of times our models chose the (correct) null model with no change-points for 500 simulated data sets of a particular sample size ($n_{\text{samp}} \in \{100, 200\}$), hazard ($\lambda_{\text{True}} \in \{0.25, 0.5, 0.75\}$) and degree of censoring (0% or 50%) within the observable time horizon. Although 0% censoring does not occur in practical applications, it serves as a reference case for assessing the impact of censoring. A value of 50% censoring is plausible, especially in trials where long-term survival is likely. Examples of datasets with approximately 50% censoring include the E1690 and E1684 datasets presented in Section 5.5 in Chapter 5.

For the scenarios with one change-point we simulated datasets with increasing and decreasing hazards, varying the difference in the hazard between intervals, while for two change-point models we also considered bathtub and inverted bathtub hazards. In these scenarios the sample size was one of $n_{\text{samp}} \in \{200, 300, 500, 1000\}$, with the true change-point equal to 0.5 and for two change-point models a change-point at time equal to 1.

A study follow-up of 2 years was assumed and observations with a survival time greater than this were censored. As noted above, for some simulation studies we assessed the impact of censoring within the study. The censoring percentages refer to the expected proportion of events within the study follow up which are censored. If a censoring percentage of 50% was required, censoring and event times were generated for 100% (i.e. double the required percentage) of the dataset with the censored time following the same piecewise distribution as the event times. This ensured that the censoring of the events occurred with approximately equal probability throughout the study follow up.

Also presented are a number of simulation study results comparing the power of our approach referred to as "Collapsing" with the model of Chapple et al. (2020), referred to as "RJMCMC".

3.6.2 Assessment of Accuracy of Survival Extrapolation

We evaluate the model's performance in estimating the Restricted Mean Survival Time (RMST) when an incorrect number of change-points was specified, compared to when the correct change-point model was selected. RMST is defined as the integral of the survival function up to a specified timepoint (Royston and Parmar, 2013).

In contrast to the simulation study assessing the power of the method to determine the number of change-points, the number of change-points of the PEM was fixed in advance, however, time maximum observed time was again 2 years with all examples assuming no censoring. The relative percentage error in RMST (Err_{RMST}) was calculated as the absolute difference between the RMST of the PEM (RMST_{PEM}) and the true RMST of the data generating process ($\text{RMST}_{\text{True}}$) divided by true RMST of the data generating process; $\text{Err}_{\text{RMST}} = \frac{|\text{RMST}_{\text{PEM}} - \text{RMST}_{\text{True}}|}{\text{RMST}_{\text{True}}}$. The RMST was evaluated up to time equal to 15 at which the survival probability would be negligible.

When the true data generating process was a constant hazard (i.e. no change-point) we calculated the Err_{RMST} when correctly assuming no change-point versus assuming one change-point respectively. In the scenarios where the data generating process had one or two change-points we compared the true change-point model versus a model with one fewer and one more change-point than the true model. The no change-point examples the simulated datasets had a sample size $n_{\text{samp}} = 100$ while in the change-point examples $n_{\text{samp}} = 300$.

3.6.3 Simulation Study Results

We tested the proposed method's ability to detect the absence of a change-point. Table 3.2 shows that the collapsing model selects the null model approximately 95% of the time, irrespective of the sample size, hazard or censoring.

Table 3.2: Power test for the no change-point model

True hazard	Probability Correct (%)	n	Censoring (%)
0.25	95.0	100	0
	96.8	200	50
	95.0	100	50
0.5	95.2	100	0
	95.8	200	50
	93.4	100	50
0.75	95.8	100	0
	96.0	200	50
	94.4	100	50

Results for one and two change-point models are reported in Table 3.3 with the same results represented pictorially in Figures 3.3 and 3.4.

For each scenario (i.e., combination of investigated parameters), we fitted the collapsing change-point model to 500 simulated datasets. Among these 500 estimated models, we identified the most probable change-point model. The frequency at which the correct number of change-points was identified is presented in Table 3.3 as % Correct. Within this subset of models that correctly identified the model, we report the average value of τ_{Est}, which represents the posterior mean of the change-point(s) (with numbered subscripts denoting the first or second change-point for two-change-point models). The associated standard error is presented in parentheses. In Table 3.3, these values are denoted as $E[\tau_{\text{Est}}]$. Also reported are λ_{True}, the simulated hazards for each interval. For clarity of exposition, we omit the expected posterior mean of the hazards and its standard error, noting that the accuracy of hazard estimation is determined by the accuracy of the change-point locations.

For the one and two change-point simulations studies, large sample sizes and/or large changes in hazards resulted in the correct model being selected with a high probability. When changes in hazards are relatively large, the correct model is selected with high probability at all samples, while for smaller changes moderate to large samples are required. Similarly, $E[\tau_{\text{Est}}]$ is closer to the true values of the change-point(s) and has a smaller standard error when there are large differences between the hazards and/or large sample sizes.

Table 3.3: Probability of selecting correct change-point model and estimation of τ_{True} with the collapsing approach

Model	Parameters	$n_{\text{samp}} = 200$	$n_{\text{samp}} = 500$	n_{sim} =1000	λ_{True}	$E[\tau_{\text{True}}]$
Increasing Small	$E[\tau_{\text{Est}}]$	0.60 (0.15)	0.56 (0.12)	0.52 (0.07)	0.5,0.75	0.5
	% Correct	38	77	97		
Increasing Large	$E[\tau_{\text{Est}}]$	0.52 (0.04)	0.51 (0.02)	0.5 (0.01)	0.25,0.75	0.5
	% Correct	91	96	97		
Decreasing Small	$E[\tau_{\text{Est}}]$	0.52 (0.14)	0.52 (0.11)	0.51 (0.08)	0.75,0.5	0.5
	% Correct	44	76	97		
Decreasing Large	$E[\tau_{\text{Est}}]$	0.49 (0.05)	0.49 (0.03)	0.5 (0.01)	0.75,0.25	0.5
	% Correct	95	96	98		
Increasing	$E[\tau_{\text{Est1}}]$	0.56 (0.1)	0.52 (0.06)	0.51 (0.03)	0.25,0.5,0.75	0.5,1
	$E[\tau_{\text{Est2}}]$	1.19 (0.14)	1.11 (0.13)	1.03 (0.08)		
	% Correct	22	57	94		
Decreasing	$E[\tau_{\text{Est1}}]$	0.34 (0.1)	0.42 (0.09)	0.47 (0.06)	0.75,0.5,0.25	0.5,1
	$E[\tau_{\text{Est2}}]$	0.96 (0.1)	1.01 (0.09)	1 (0.05)		
	% Correct	20	47	90		
Bathtub	$E[\tau_{\text{Est1}}]$	0.48 (0.03)	0.49 (0.02)	0.5 (0.01)	0.75,0.2,0.75	0.5,1
	$E[\tau_{\text{Est2}}]$	1.02 (0.04)	1.01 (0.02)	1 (0.01)		
	% Correct	89	94	95		
Invert Bathtub	$E[\tau_{\text{Est1}}]$	0.51 (0.03)	0.5 (0.01)	0.5 (0.01)	0.2,0.75,0.2	0.5,1
	$E[\tau_{\text{Est2}}]$	0.99 (0.03)	0.99 (0.03)	1 (0.01)		
	% Correct	92	95	97		

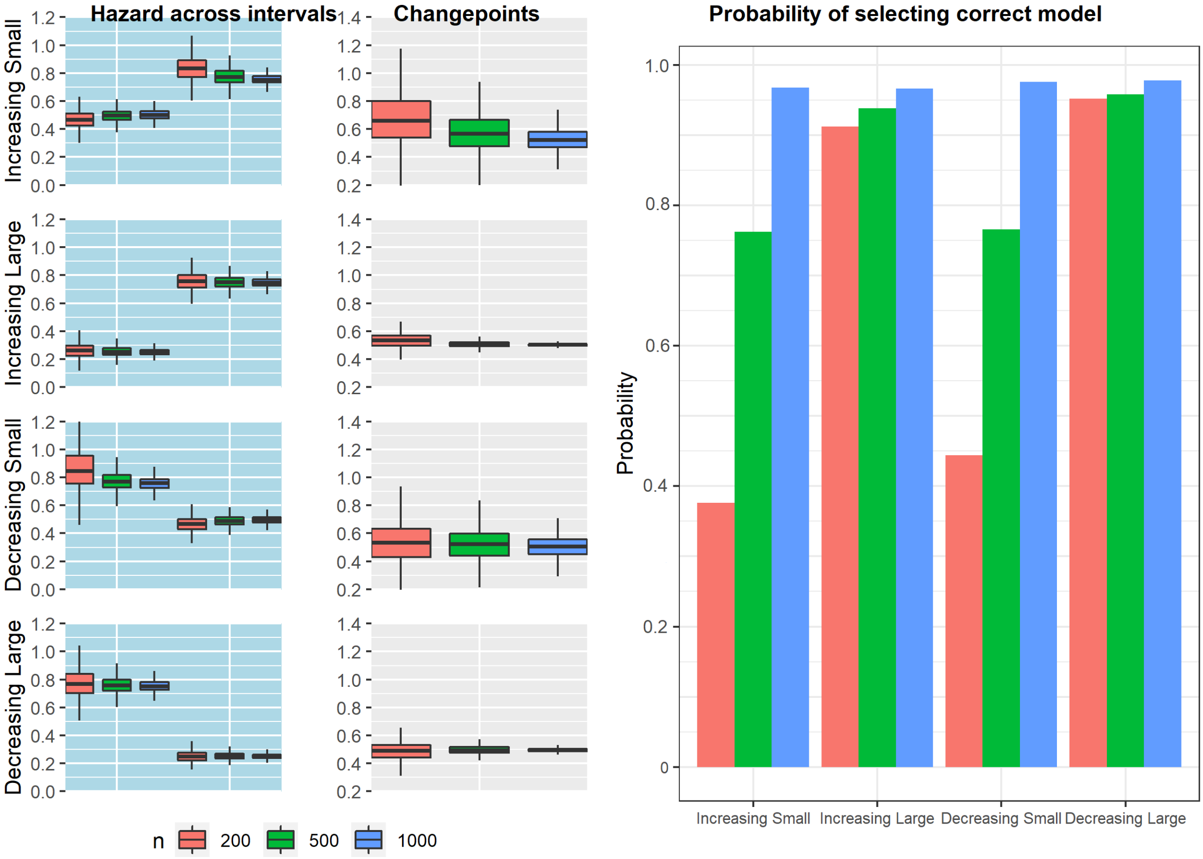


Figure 3.3: Simulation study results for the 1 change-point scenario

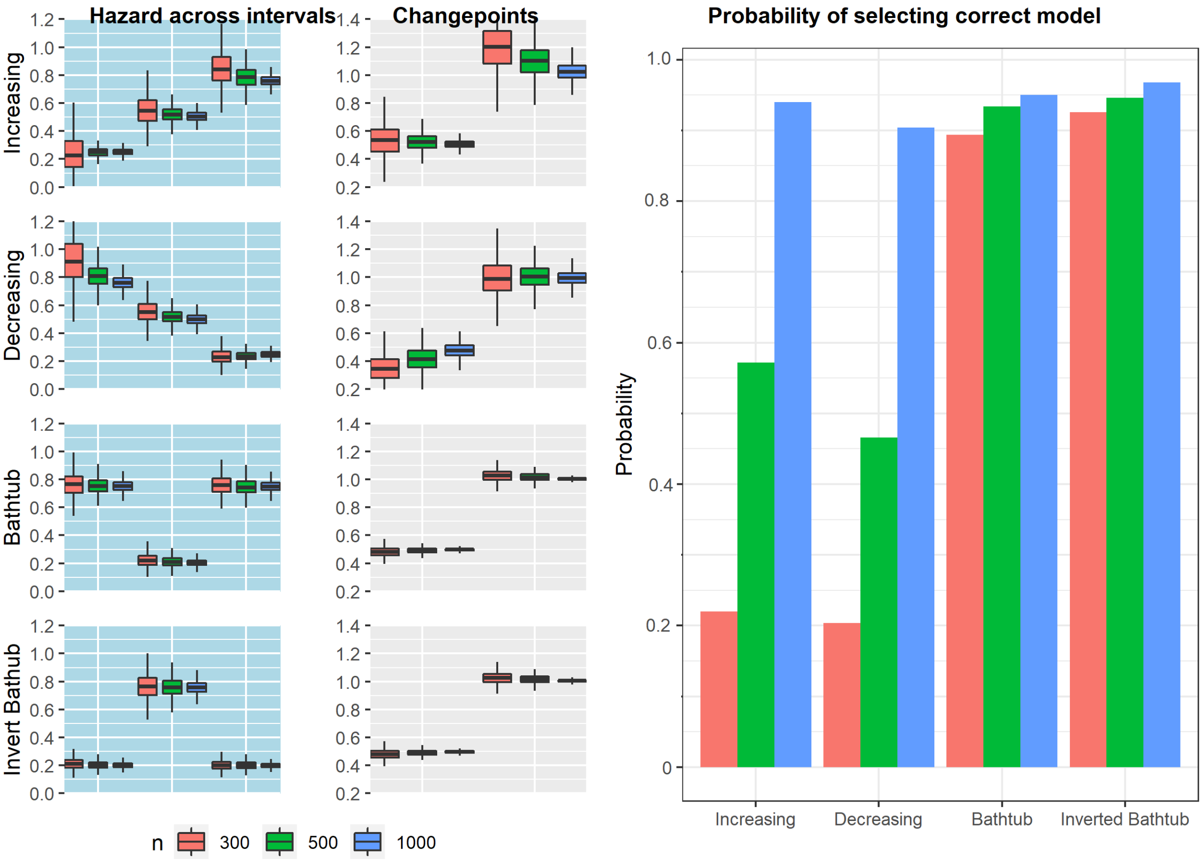


Figure 3.4: Simulation study results for the 2 change-point scenario

Considering the comparison between Chapple et al. (2020) and our approach, for each of the one and two change-point scenarios we found that the collapsing approach had a higher probability of detecting the correct change-point model as illustrated in Figures 3.5 and 3.6.

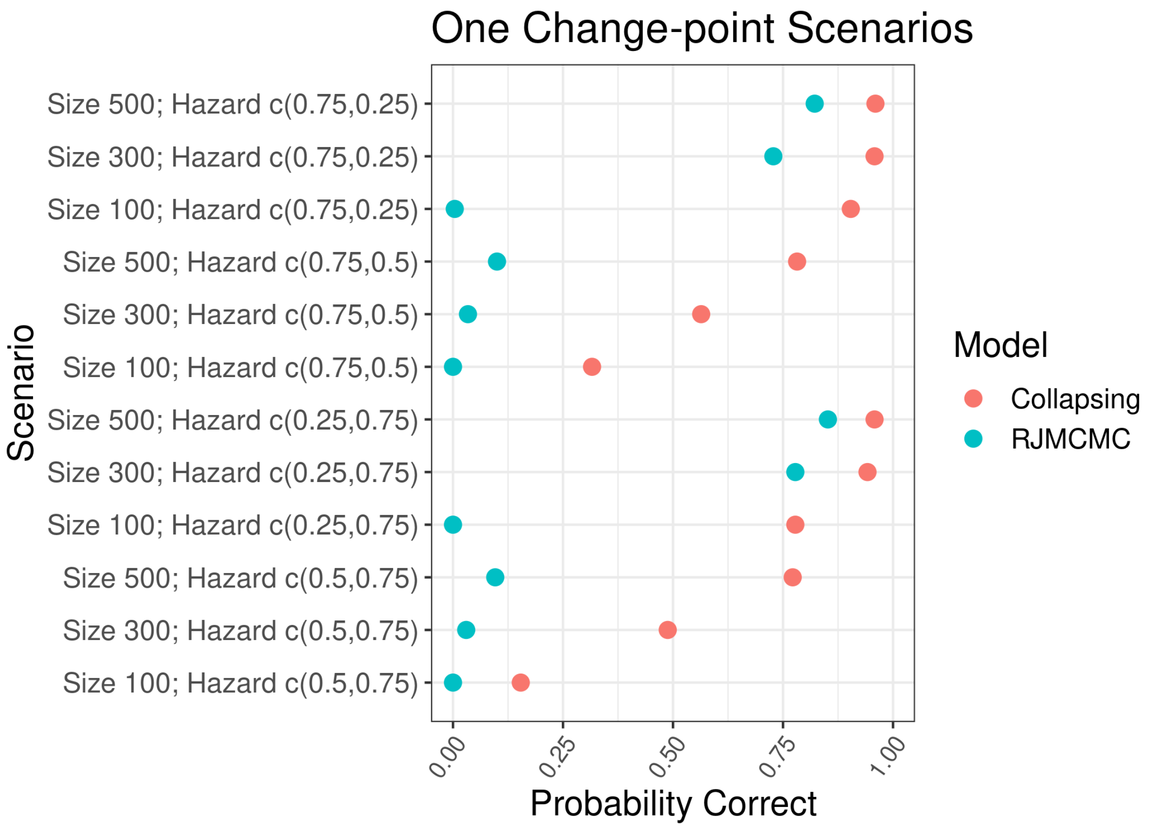


Figure 3.5: Probability of correctly detecting the change-point (1 change-point scenarios)

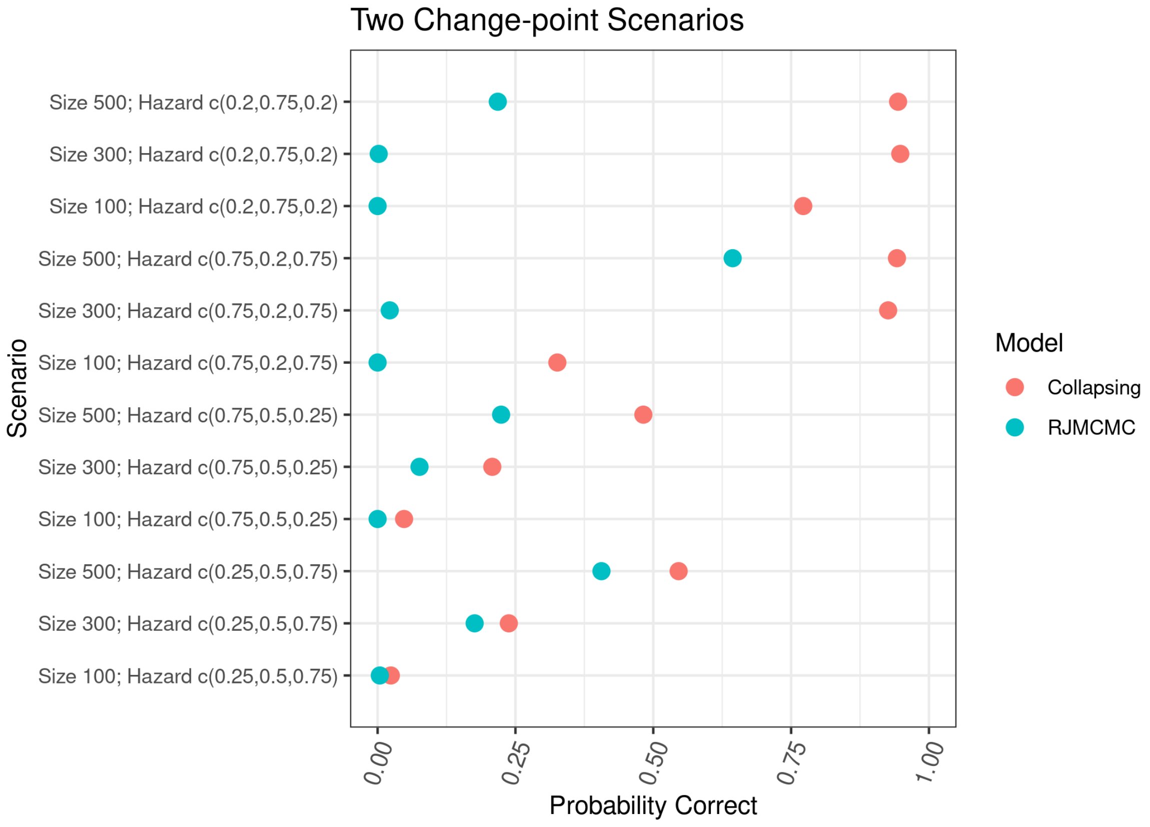


Figure 3.6: Probability of correctly detecting the change-point (2 change-point scenarios)

Results for the scenarios predicting long-term survival are presented in Table 3.4. As expected, the percentage error is reduced when there is a lower survival probability at the end of follow-up (i.e. models with higher initial hazards). Additionally, the error increases when the model is incorrectly specified, however, the increase is quite modest, particularly when an extra change-point is estimated. As noted previously, the prior for the change-point places lower prior probability on change-point locations where a change-point is placed close to another change-point or last observation. This means that the final interval which informs the extrapolated hazard has a relatively large sample size increasing the robustness of the extrapolations.

Table 3.4: Accuracy in estimating Restricted Mean Survival with different change-point numbers

	$\text{Err}_{\text{RMST}}\%$		
Model	Correct	Incorrect - One too few	Incorrect - One too many
Constant ($\lambda = 0.25$)	8.34	-	11.07
Constant ($\lambda = 0.5$)	5.61	-	7.03
Constant ($\lambda = 0.75$)	3.75	-	4.85
Increasing Small	2.53	4.16	3.03
Increasing Large	2.27	8.4	2.82
Decreasing Small	4.21	6.24	4.97
Decreasing Large	6.01	22.02	6.61
Increasing	2.67	2.87	3.31
Decreasing	7.98	8.46	8.1
Bathtub	2.55	4.8	2.84
Invert Bathtub	7.37	14.56	7.66

3.7 Applications

In this section we applied the approach to real data sets. In our first application, we investigate how the method can be used to explore the behaviour of the hazard. In our second example we assess the performance of the change-point model in comparison with several popular survival models in the context of survival extrapolation.

3.7.1 Leukaemia Remission times

Matthews and Farewell (1982) present times from remission induction to relapse for 84 patients with acute nonlymphoblastic leukaemia who were treated on a common protocol at university and private institutions in America (see Glucksberg et al. (1981)). The author noted that this data set is typical of those encountered in the treatment of acute leukaemia, except that 24 of 33 censored observations were censored at six months (182 days) when the patients were randomized to an experimental protocol. Of the 84 patients, 51 patients had an event, with the median time-to-event being 284 days, and survival times ranging from 24 days to a patient who was censored at 2057 days. One topic of clinical interest was determining if a treatment provided a reduction in the hazard of relapse after a period of time, relative to the initial time after induction, which was typically assumed to have a high constant hazard of relapse.

Matthews and Farewell (1982) considered the detection of a change-point in the hazard of relapse for leukaemia using a frequentist approach. Using a log-likelihood ratio test the null hypothesis of no-change-point was rejected. It is however, possible that there are additional change-points in the data. To be consistent with the analysis of Matthews and Farewell (1982), we removed observations which are censored at 182 days.

Using the Gibbs sampler detailed in Section 3.4.1, change-point models with 0-4 change-points were fit to the data. Based on diagnostic statistics of pilot runs from Gelman and Rubin (1992), Raftery and Lewis (1992) and Geweke (1991) each model was run for 10,100 simulations, with the first 100 discarded. The log mean marginal likelihoods and corresponding Bayes Factors for the models are presented in Table 3.5. Using the criteria presented in Table 3.1 there is clear evidence supporting the one change-point model versus the no-change-point model. Considering model fit using Bayes Factors, although the models with 3 or more change-points are more probable than the one change-point model, they do not offer sufficient evidence to favour them over the simpler one change-point model. When the data was analyzed with the collapsed change-point model, the 2 change-point model was estimated as being slightly more probable than the one change-point model, however, there was a substantial posterior probability for the one change-point model. Models with higher number of change-points had additional change-points at earlier timepoints, however, the posterior distribution for final change-point is very similar across all models. This suggests that the final change-point is at approximately 642 days, however, additional change-points in the early portion of the data may provide a slightly better fit to the observed data.

Based on this analysis we propose that the hazards experienced by these leukaemia patients is best described by a one change-point model. Figure 3.7 shows the mean posterior change-points and survival probabilities (green diamond and purple line) along with the 95% credible interval for survival (dashed grey line). Table 3.6 confirms that the hazard falls considerably after 642 days (1.76 years).

Table 3.5: Marginal Likelihood of change-point models applied to leukaemia remission data

Model	Log_{10} Mean Marginal Likelihood	Bayes Factor*	Posterior Probability†
0	-161.22		1.9 %
1	-159.81	25.4	38.2 %
2	-159.45	2.33	38.6 %
3	-159.29	1.43	16.8 %
4	-159.18	1.28	4.5 %

∗ Bayes Factor versus previous model

† Collapsing change-point Model

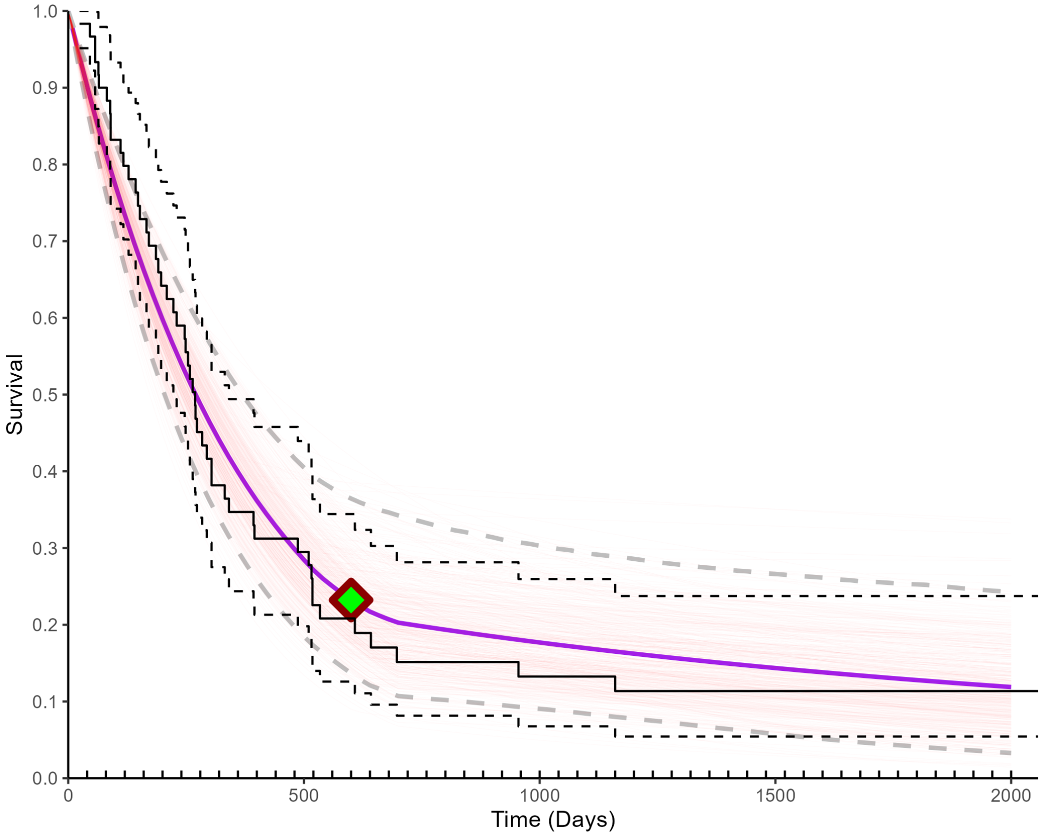


Figure 3.7: Survival function of one change-point model applied to leukaemia remission data.

Table 3.6: Posterior Quantiles for the one change-point model parameters.

Parameter	Median [95% Credible Interval]
τ_1 (days)	642 [304-697]
λ_1 (per year)	0.936 [0.69 - 1.1235]
λ_2 (per year)	0.143 [0.021 - 0.473]

Figure 3.8 shows the posterior probabilities of the change-points and the maximum *a*

posteriori (MAP) estimate of the change-point is 697 days, which is consistent with the results from Kim et al. (2020) and Matthews and Farewell (1982) who obtained a maximum likelihood estimate of 697 days.

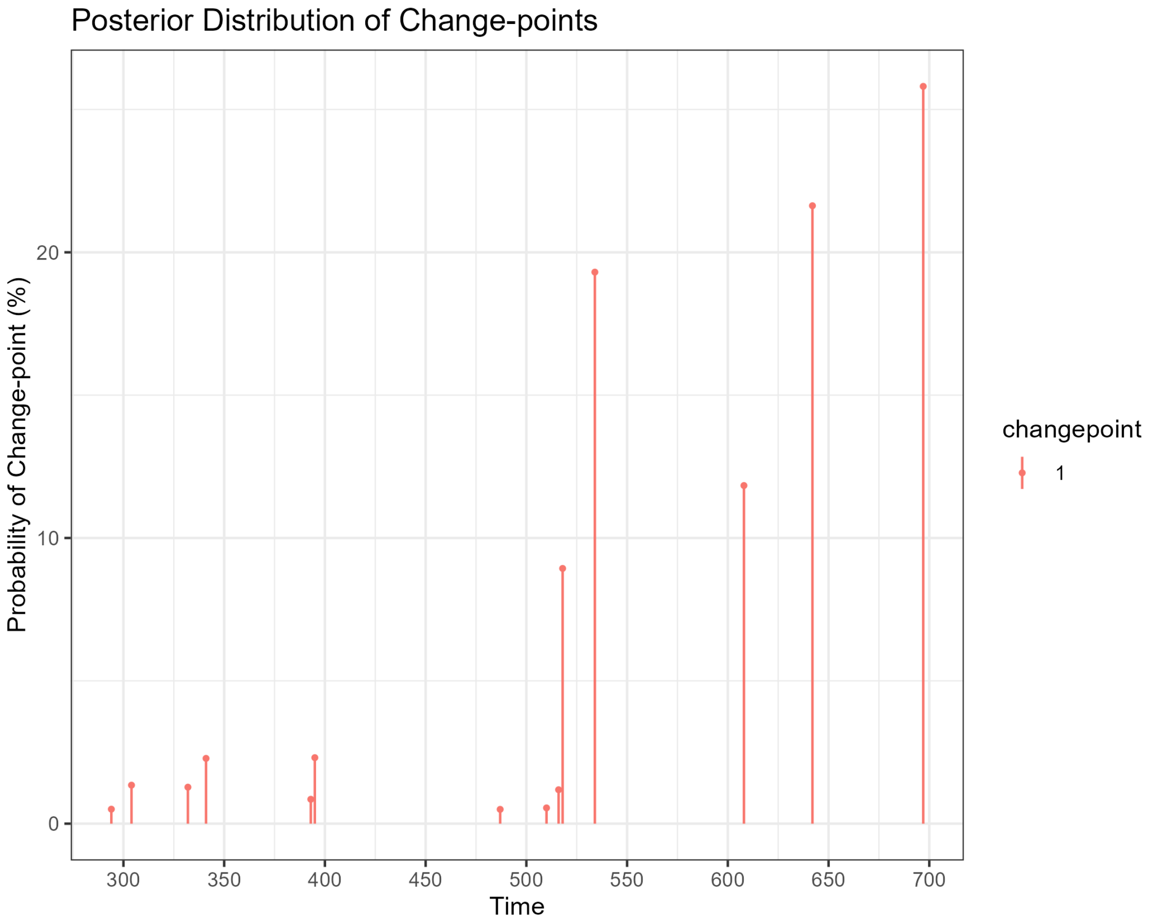


Figure 3.8: Posterior probabilities of change-point locations for one change-point model applied to leukaemia remission data.

3.7.2 Glioblastoma data: Identifying trends in the hazard function

One potential application of hazard change-point analysis is the visualisation of the hazard function itself. We compare the hazard function estimated from the Gibbs sampler with approaches documented by Hagar and Dukic (2015), who review a variety of packages used to estimate the hazard in time-to-event data for the statistical software R.

We consider data relating to survival times for Glioblastoma, a central nervous system cancer in which prognosis remains poor. The data is available using the R package `RTCGA` developed by Kosinski and Biecek (2020) and contains a sample of 595 patients of which 446 experience death. The median survival is 1 year with approximately 5% of patients surviving after 5 years.

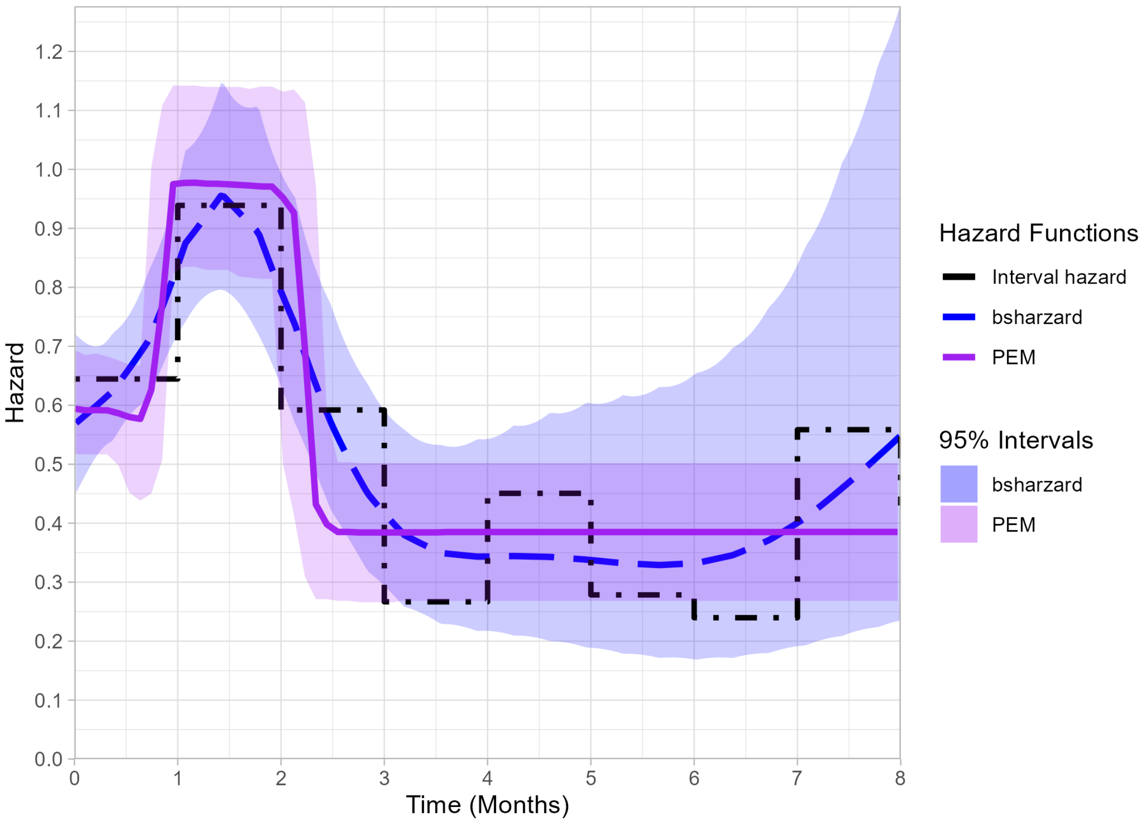


Figure 3.9: Step (black-twodash), Smoothed (blue-longdash) and posterior mean (purple-solid) hazard functions applied to Glioblastoma data.

Figure 3.9 below provides an estimate of the hazard of death for the Glioblastoma data using three approaches. The first approach, coloured in black (twodash line), divides the time interval into bins of equal width (in this case one-year intervals), and then estimates the hazard in each bin as the number of events d_i in that bin divided by the number of patients at risk in each interval, n_i with the hazard for that interval $h_i = \frac{d_i}{n_i}$ (see R package `muhaz` by Hess and Gentleman (2019)). The second approach uses B-splines from a generalized linear model perspective to estimate a smoothed hazard function along with confidence regions (see R package `bshazard` by Rebora et al. (2018)), coloured in blue (longdash line) with confidence regions in grey. The third approach, plots the posterior median of the hazard function.

From Figure 3.9 it appears that the hazard peaks at 1-2 years and then falls thereafter. The posterior distribution of the change-points are concentrated at times 0.85 and 2.25 and their 95% credible intervals do not overlap. The posterior distributions of the hazards also do not overlap with the posterior distribution of an adjacent interval, suggesting a clear change in the hazards between each interval. Using the posterior medians of the parameters we can surmise that there are three distinct intervals; in the first interval up to approximately 0.85 years, there is a moderately large hazard of approximately 0.6. Then from the period 0.85 to 2.2 years the hazard peaks around 1 and falls to approximately

0.4 thereafter. The finding that the hazard is peaked is consistent with Wang et al. (2015), however, the long-term drop in hazards is more pronounced for patients in this dataset.

3.7.3 Predicting survival by extrapolating constant hazards

Miller and Halpern (1982) presented survival times for 184 patients who received heart transplants. Visual inspection of the cumulative hazard plot suggests that after 1 year the hazards are approximately constant (i.e. linear). Assuming this to be correct, we artificially censored the data at 2 years and fit the piecewise exponential model and other commonly used survival distributions to this data estimated with the JAGS and Stan software programs (Plummer, 2003; Stan Development Team, 2020). For each model we assessed the statistical fit to the partially observed data and the difference in predicted survival to the fully observed data.

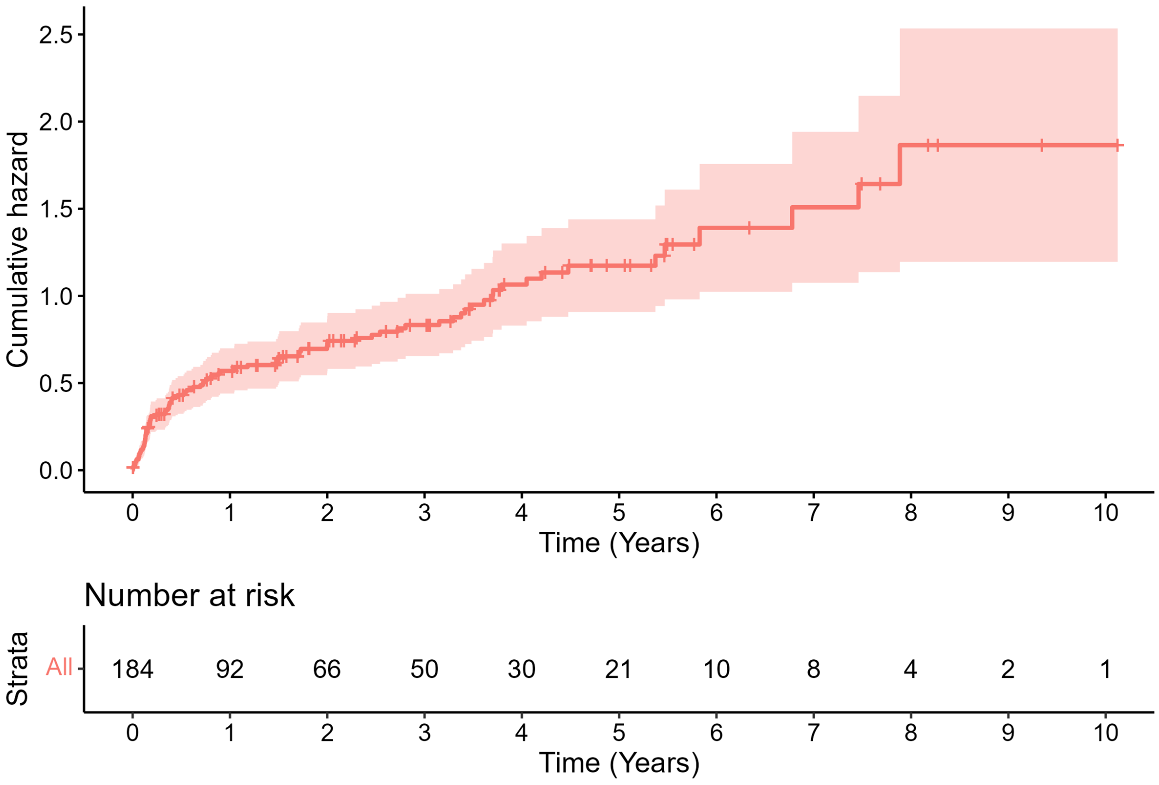


Figure 3.10: Cumulative Hazard Plot for Stanford Heart Data

Statistical fit was assessed through a number of criterions, namely Pseudo-Marginal Likelihood (PML) and Widely Applicable Information Criterion (WAIC) with details on their respective computation available in A. E. Gelfand (1994) and Watanabe (2010).

Similar to the previous section we "uncollapse" the hazards at each simulation and calculate the survival function with the mean posterior survival being the average of these survival probabilities at each timepoint. For the piecewise exponential model we found that the 2 change-point model had the highest posterior probability $\approx 66\%$. The posterior mean of the first change-point is 0.18 years at which the hazard falls from a posterior mean of 1.56 to 0.42. The posterior mean of the second change-point was 0.81 years after which the posterior mean of the hazard was 0.16. Figure 3.10 highlights the piecewise exponential model provides a predicted survival similar to the long-term Kaplan-Meier survival. In Figure 3.10 green diamonds refer to the mean change-point locations for the change-point model. Dashed line refers to the timepoint at which the original dataset was artificially censored.

Table 3.7 presents the PML and WAIC (evaluated using the R package loo by Vehtari et al. (2020)) for each of the models fit to the partially observed data. The table also presents the RMST evaluated up to 10 years based on the expected extrapolated survival for each model. The RMST estimated from the (long-term) Kaplan-Meier survival function (i.e., estimated using the data which were not artificially censored) is also presented. Because the Kaplan-Meier survival function is an estimator of survival and is subject to statistical uncertainty, the uncertainty in the survival function was estimated using a bootstrapping procedure. The RMST was calculated for each bootstrapped replication and then averaged. To estimate how closely the parametric survival models predicted the true long-term survival data, we calculated the expected absolute difference (in years) between the RMST estimated from the bootstrapped replicated Kaplan-Meier survival function and the RMST for the parametric model ($E[|\text{RMST}_{KM} - \text{RMST}_{Model}|]$). Consistent with the hypothesis that the long-term hazards were approximately constant, the piecewise exponential approach is the best fit to the true data, both in terms of statistical fit and deviation in terms of RMST.

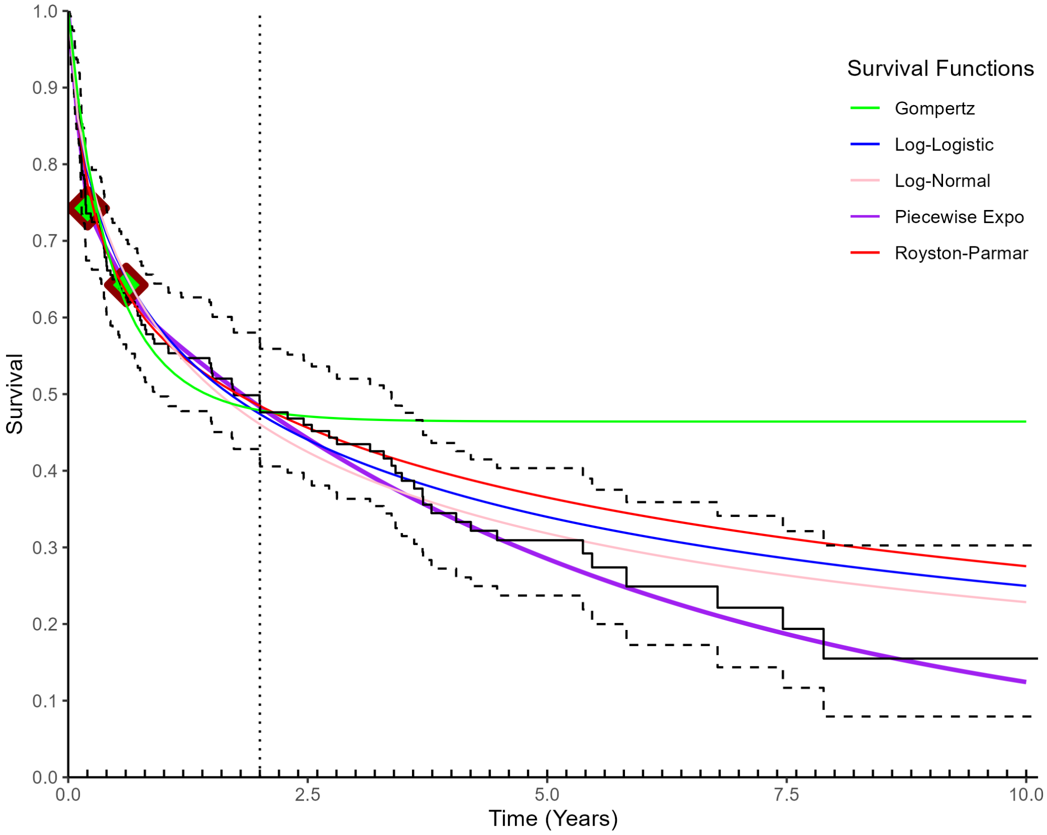


Figure 3.11: Long-term survival probabilities for 5 best fitting survival models.

Table 3.7: Goodness of fit statistics for survival models and comparison of RMST between models and the long-term data

Model	-2log(PML)*	WAIC	RMST	$E[\lvert \text{RMST}_{KM} - \text{RMST}_{Model} \rvert]$
Piecewise Exponential	250.88	250.8	3.31	0.37
Royston-Parmar 1 knot	264.17	264.16	4.04	0.67
Log-Normal	265.72	265.72	3.86	0.54
Log-Logistic	268.95	268.95	3.68	0.47
Gompertz	269.61	269.61	4.91	1.46
Generalized Gamma	271	271	3.52	0.4
Weibull	274.53	274.53	3.38	0.4
Gamma	278.88	278.88	3.15	0.49
Exponential	321.32	321.32	2.21	1.45
True Observations			3.44	

* -2log(PML) was calculated to place it on the same scale as WAIC. Lower values indicate better fit.

3.8 Discussion

In this chapter we have presented two Bayesian approach to determining the number and location of change-points in a hazard function including the special case were no change-point exists. By employing a Bayesian approach, uncertainty around the number of change-points is automatically computed and is described with a probabilistic interpretation, allowing us to assess the relative evidence of alternative change-point models. In the Gibbs sampler approach we fit all candidate change-point models (including the null model) and select the change-point model using Bayes Factors and the decision rules suggested by Jeffreys (1961).

In the collapsing model approach we take advantage of the fact that the marginal likelihood for a piecewise exponential model without covariates can be expressed analytically. By restricting the change-points to be event times we reduce the complexity of the parameter space resulting in a simpler and computationally efficient algorithm. While the approach of Chapple et al. (2020) can be considered more general in that it allows covariates and continuous change-points, we found that in some examples the change-points were highly correlated due to the relative infrequency of moves between model dimensions and that for smaller changes in the hazard change-points were not detected. Furthermore, as demonstrated in our simulation study the collapsed approach had a higher probability of detecting the correct change-point model.

It should be noted that evaluating the marginal likelihood and subsequent model selection using Bayes Factors from the Gibbs sampler model is similar to the collapsing model which visits the competing change-point models proportional to their marginal likelihood. For the collapsing change-point model the marginal-likelihood is also multiplied by a prior for the change-point number (i.e. the Poisson prior), while for the Gibbs sampler the model selection is based on the marginal likelihood and the decision rule suggested by Jeffreys.

Our simulation studies demonstrated the ability of the algorithm to detect change-points when sample size and/or change in hazards is large along with the consistency of the estimators. As demonstrated by the no change-point simulation study, the model has a low probability of detecting the presence of change-points when they do not exist. As with any Bayesian analysis, the inferences are somewhat informed by the choice of prior. We consider a discrete prior on the change-point locations which has the advantage of ensuring that the change-points are not too close together or close to the final events, where there is typically a sparsity of data. Because model selection is based on the evaluation of the marginal likelihood, this calculation can be sensitive to the hyperparameters used, and we provide an approach to specify a hyperprior for the β hyperparameter which we describe in Appendix A.1.1. The Poisson prior on the change-point number is reasonably robust to alternative specifications, however, in individual examples where posterior model probabilities are similar, it will naturally have some effect.

In this chapter we have presented real world applications of piecewise change-point models. For the leukaemia data we considered the relative probability of multiple change-point models rather than assuming the choice was between a one or no change-point model. Regarding the Glioblastoma data, our approach segments the hazard function into distinct intervals which may allow greater interpretability of the trends in the hazard function even when not considering the piecewise exponential model for survival extrapolation. In situations where the constant long-term hazards are plausible we believe that a piecewise exponential model should be considered and in the third example we compare the extrapolated survival of the piecewise exponential model vs that of other parametric models. Although we artificially censored the data for the purpose of our example it is reasonable to hypothesize that these heart transplant patients may be subject to different hazards as time progresses. Patients are likely subject to high hazards of death during or immediately after a complex surgical procedure such as a heart transplant. Over the initial number of months patients are likely to be at an elevated risk of transplant rejection and many events may occur over this period. If patients do not reject their transplanted organ, over the long-term they are subject to a lower hazard associated with all-cause mortality. When considering the lifetime time horizon, piecewise exponential models (and any parametric model) should always be adjusted to ensure the extrapolated hazards do not fall below general population mortality which we discuss in greater detail in the next chapter.

We note the relatively recent Technical Support Document (TSD) regarding flexible survival models (Rutherford et al., 2020). Regarding the use piecewise exponential models in health technology assessment, they state that "the cut-points for the various intervals may be arbitrary and may importantly influence the results of an analysis" and "splitting the data into sections according to time means that sample sizes are reduced in later segments of the curve". We believe our approach addresses both of these limitations as firstly the location (and number) of change-points is informed by the data and secondly the prior we use for the change-points reduces the probability that change-points close together or to the final event will be selected. Rutherford et al. (2020) also highlight situations in which the Kaplan-Meier survival function is used to represent the initial section of the survival function and an exponential function is adjoined to a predetermined point of the Kaplan-Meier. In this situation, our approach could also be used in determining the final change-point and its associated uncertainty from which the constant hazard is extrapolated. In order to test this hypothesis, in the subsequent chapter we review previous applications of Bagust & Beale approach to HTAs and apply the collapsing model to the survival outcomes.

4 Piecewise Survival models in HTA

4.1 Introduction

As described in Chapter 2, there are a number of methodologies to extrapolate time-to-event outcomes. One framework developed by Bagust and Beale (2014) suggest examining the clinical trial for biologically plausible hypotheses rather than assuming the data are best described by one of the standard parametric models. One key recommendation of this approach is that the trial protocol (or disease processes) may induce transient effects on the hazard of an event. Bagust and Beale suggest visually inspecting the empirical cumulative hazard function to identify a timepoint after which there is evidence of a long-term linear trend. Once identified, a constant hazard (exponential) model should be fit to the data after this timepoint.

As highlighted in Section 2.5.2 there are several important issues are associated with the B&B approach particularly relating to the selection of the timepoint and that the non-parametric nature precludes comparison with standard parametric models.

We propose to use the statistical method described in Chapter 3 which addresses these limitations. This method is fully parametric and sits within the framework of NICE TSD 14. It objectively identifies the timepoint after which a constant hazard appears plausible. We consider the application of the B&B approach in previous HTAs and compare the concordance of the timepoint estimated by the B&B approach to the proposed method. In situations where more mature survival data became available after that which was presented in the HTA, we assessed the accuracy of the extrapolated survival using the proposed method compared to other parametric models. While the HTAs considered are those submitted to NICE, the method described is of relevance to a wide range of HTA authorities whose assessments require extrapolating survival outcomes.

4.2 Methods

4.2.1 Identification of submissions using Bagust and Beale approach to extrapolate survival

We identified HTAs which used the B&B approach based on a previous review by Bell Gorrod et al. (2019). The scope of the review was restricted to completed NICE single technology appraisals (TAs) for cancer treatments that commenced between July 1, 2011, and June 30, 2017. A TA involves a company submission providing evidence on a single technology for a single indication for the purpose of making recommendations for use in the National Health Service (NHS) for England. During the process, the evidence is assessed by an independent Evidence Review Group (ERG), which produces its own report and conclusions. The TAs which were screened were identified as those categorized as "piecewise" in the Bell Gorrod et al. (2019) review. For each of these TAs one of the authors assessed if the B&B approach was employed for extrapolating survival outcomes. If the B&B approach was used, information on the timepoint after which constant hazards were assumed and Kaplan-Meier (KM) survival functions were extracted. The selection of relevant TAs and accuracy of extracted data was confirmed and verified by the second author. A three-step process was then performed for each identified KM survival function. First, the KM survival function was digitized using the WebPlotDigitizer application (Rohatgi, 2022). Individual patient data was then reproduced using the algorithm by Guyot et al. (2012) using R version 4.1 (R Development Core Team, 2023). KM survival functions estimated from the reconstructed patient data were inspected by both authors to assess agreement with the original KM survival function. Information on each of the TAs considered in our review and the relevant information extracted from them are detailed within Appendix B.2.

4.2.2 Alternative to Bagust and Beale approach using a change-point model

The key step in the B&B approach is the identification of a point at which the hazard changes and remains relatively constant. Rather than seek to identify this timepoint subjectively, we estimate this statistically, using a change-point model. A change-point model assumes that the data generating process can undergo changes over time such that one model will not be appropriate for all time periods. The times at which the statistical model undergoes abrupt changes are termed the change-points. These split the data into contiguous segments. The parametric model assumed for the data usually remains the same across segments (e.g. exponential model), but changes occur in its model parameters. Change-point models can also be called "piecewise models/approaches", however, this term has also been used to describe situations in which some portion of the trial/external data are modelled non-parametrically (i.e. with a KM survival function) (Bell Gorrod et al., 2019; Latimer, 2013; Rutherford et al., 2020). Therefore change-point models are more accurately described as a particular type of piecewise model. Specifically, we consider a particular change-point model called a piecewise exponential model (PEM), a survival model which assumes that the hazard function is constant within segments but independent between segments. The PEM, which we described mathematically in Chapter 3, uses the data directly to estimate the number and location of change-points. The final segment of the data estimates the constant hazard used to extrapolate survival. As stated in Chapter 3 the PEM tends to avoid overfitting as we use a prior belief on the change-point locations which has the advantage of ensuring that the change-points are not too close together or close to the final events, where there is typically a sparsity of data. By using a Bayesian approach, the posterior distribution of the change-point(s) and their locations automatically characterize parameter uncertainty and can propagate this uncertainty into the survival function. Also produced are the relative probabilities of each change-point model (i.e. none, one, two change-points) conditional on the observed data, so that the survival function can be either based on the most probable change-point model or a weighted average (with weights based on relative probabilities) of all PEMs. The exponential model is a special case of the PEM, thus we avoid the issue of identifying a subset of the data as having a constant hazard when the entire data can be adequately described with a single constant hazard.

As noted in Chapter 3 we verified this property by conducting simulation studies which show the power to detect a constant hazard model was above 90%. Simulation studies also show that the capability of the method to estimate the true model parameters increases with the sample size. Of particular interest to this use case, the relative error in predicting extrapolated survival when the incorrect number of change-points was chosen vs when the correct number of change-points was used was also estimated. Estimating additional change-points resulted in a lower increase in the error associated with extrapolated survival than assuming a model with too few change-points.

Because the number and location of the change-point(s) are determined by the data, the PEM can provide a good fit to the data even in situations where there are large decreases or plateaus in the survival function, which can sometimes be observed at the beginning of the trial (Bagust and Beale, 2014). The model is fully parametric and can be compared to other models using information criteria such as the widely applicable information criterion (WAIC) (Watanabe, 2010). The method is implemented as an R package with information on its installation along with a guide on its use in Appendix 8.1. Also included is a Visual Basic Function so that a user can easily calculate the survival of a PEM in Microsoft Excel given the change-points and segment hazards. The assumption of constant hazard for long-term extrapolation is very strong, especially if a proportion of the trial survive to an age where general population mortality is no longer negligible relative to the hazard observed in the trial. It could be plausible, however, that once general population mortality (GPM) has been accounted for, that disease-related hazards can be relatively constant, with PEMs previously used in this context (Estève et al., 1990). For a wide range of diseases the Surveillance, Epidemiology, and End Results (SEER) Program (2023) database provides tools to estimate the annual cancer specific conditional survival. If the conditional survival probabilities become relatively similar across the time intervals, it could provide clinical plausibility for (approximately) constant disease specific hazards. Within the short timeframe of a typical randomized control trial, the disease specific hazards are often similar to the all-cause mortality rate (Rutherford et al., 2020), and one possibility is to add GPM hazards to the hazards predicted by the parametric model beyond the follow-up of the trial (externally additive hazards). A variety of other approaches are described elsewhere by van Oostrum et al. (2021).

4.2.3 Analysis of reconstructed datasets using Piecewise Exponential Model

PEMs were fit to the reconstructed datasets which were previously analysed using the B&B approach with further information provided in Appendix B.2. As part of model estimation, the relative probabilities for each change-point model (in terms of number of change-points) were estimated. We reported the mean value of the final change-point along with its 95% credible interval for the change-point model with the highest probability.

4.2.4 Comparison of Extrapolated and Observed Survival from Extended Follow-up

For each TA identified using the B&B approach we identified the pivotal trial used to estimate the survival models and its associated data-cut off (DCO). We then searched for any publications which provided updated KM survival functions of the relevant survival outcomes (through manual searching and by identifying the linked publications at ClinicalTrials.gov. (2023)) which were digitized using the process previously described. Parametric survival models were fit to the digitized data derived from the KM survival functions based on the original submission. Fitted distributions included the exponential, Weibull, gamma, Gompertz, log-logistic, log-normal, generalized gamma, Royston-Parmar cubic spline and PEMs. For the Royston-Parmar models, both one and two knots were considered with the number of knots selected based on WAIC. For the PEM, the number of change-points was determined by selecting the most probable change-point model. In all cases the survival function generated from this most probable model was similar to the survival function obtained from a weighted average of PEMs. For each of the updated DCOs, updated survival outcomes were compared against the predicted survival of the PEM and other parametric models. Statistical goodness of fit was assessed using WAIC values obtained from the parametric models. We adjusted all survival functions for GPM, using the external additive hazards mentioned in the previous section with GPM data sourced from life tables of the United Kingdom, which present the annual mortality probabilities observed within the general population of the country for males and females Office for National Statistics (ONS). We considered a cohort approach, whereby for each timepoint after baseline the age of the cohort is simply the average age in the trial at baseline plus the time since the baseline. Furthermore, the proportion of males vs females is assumed to remain constant over the considered time horizon. We converted the gender-weighted mortality probabilities to hazard rates which were then added to the extrapolated hazard from the parametric model to obtain the all-cause hazards and consequently the estimated survival. We discuss other approaches to implementing general population mortality along with their potential advantages/limitations in Chapter 8.1.3. We also demonstrate in Chapter 8.1.3 that for the examples considered in this study the different approaches to including GPM will yield almost identical results. To assess accuracy of the predicted survival we calculated the RMST for all models until the maximum time in the updated DCO. Because the Kaplan-Meier survival function is subject to uncertainty, we calculated the (average) RMST based on 1,000 bootstrap replications. For each of these bootstrapped values we also calculated the absolute difference between the bootstrapped RMST (from the Kaplan-Meier) and the average RMST for each parametric model, which was then averaged.

4.3 Results

4.3.1 Previous submissions using the Bagust and Beale approach

Three of the fourteen TAs categorized as "piecewise" used the B&B approach. As a justification for their application of the B&B approach, these TAs cited three previous TAs that also used the B&B approach and were within the scope of the review. All six TAs are summarized in Table 4.1. We refer to individual TAs using the index number assigned to them by NICE. In two TAs (TA268, 2012; TA347, 2015) the B&B approach was applied by the ERG as a scenario analysis rather than the manufacturer's basecase submission. Justification for the chosen timepoint at which a constant hazard began was primarily based on visual inspection of the cumulative hazard function (TA268, 2012; TA347, 2015; TA269, 2012; TA447, 2017). A linear regression line fit to the identified interval of the cumulative hazard function was used for additional justification in TA268 (2012); TA347 (2015) and TA269 (2012). Although TA428 (2017) included cumulative hazard plots, the primary justification was based on visual fit of the survival function. A more formal statistical approach using the `piece.linear` function from the `SiZer` package (Sonderegger, 2022), was conducted in TA396 (2016). This approach assumed that the timepoint for the constant hazard extrapolation was the likelihood-maximising change-point of the cumulative hazard function. The use of regression methods and linear modelling assume that points are distributed normally around a line of best fit, which is not appropriate in survival analysis, as it does not account for censoring or that the cumulative hazard function is an increasing step function. In most cases (TA268, TA269, TA396, TA428) there was general agreement between the timepoint chosen in the TAs and the final change-point from the most probable change-point model, with the chosen timepoint either having a difference of less than 3 months compared to mean change-point value and/or being within the final change-point's 95% credible interval. An exception to this was TA347 (2015) in which an ERG assumed that for the overall survival outcome (OS) for nintedanib + docetaxel, the period after about 6 months had approximately constant hazards. Using the PEM approach, we found that the posterior probability of a no change-point model was ≈70%, suggesting that an exponential model adequately fit the observed data. For the docetaxel monotherapy arm there was evidence of a change-point, however, the mean value of the change-point from the PEM was 2.5 months compared with approximately 9 months as assumed by the ERG. For progression free survival (PFS), although not identified by the B&B approach, it was assumed that for both treatment arms there was a common constant hazard after 12 months. Fitting PEMs to the data, the most probable change-point model indicated average final change-points at about 7 and 5 months for the nintedanib + docetexal and docetaxel monotherapy arms respectively. The average hazard after the final change-point was very similar for the PEMs fitted to each of the treatment arms, supporting the assumption of a pooled extrapolated hazard. Because the OS and PFS data were quite mature (<20% survival at end of follow-up) the differences in the extrapolations between the change-point and no change-point (exponential) model were minimal. Another situation with differences in results was TA447 (2017) where the applicant assumed that the constant hazard began at 5 months. Using the PEM approach, the most probable change-point model was the no change-point model, with 80% probability. In TA531 (2018), which superseded TA447 (2017) with a later DCO, the ERG applied exponential extrapolations to both arms at ≈ 10 months. Applying the PEM to this data, the most probable model remained the no change-point model (probability of ≈ 70%). The estimated change-point locations for the one change-point model where 6 months and 12 months, both similar to the values used in TA447 and TA531 respectively. Because the survival data were relatively immature (65% and 45% survival at end of follow-up for the DCOs used in TA447 and TA531 respectively) and the hazard after the change-point was lower than before the change-point, there was a difference in predicted survival. For the one-changepoint model estimated with the DCO used in TA477, the median predicted survival at 60 months was 22.3% compared with 16.3% from the exponential model.

Table 4.1: Technology Appraisals assuming a Kaplan-Meier + constant hazard extrapolation.

						OS		PFS	
						Timepoint (months) after which constant hazard was assumed*			
Appraisal	Year of Assessment	Outcomes	Approach used in TA	Treatment	Intervention/ Comparator	TA	PEM	TA	PEM
TA268 (2012)	2012	PFS and OS both arms	Visual Inspection	Ipilimumab	Intervention	25.3	25.4 [24.6-25.6]	12	14.6 [10.3-18.7]
				Glycoprotein 100	Comparator	11.2	13.8 [9.1-25.1]	6.2	7 [6-12.4]
TA269 (2012)	2012	PFS both arms		Vemurafenib	Intervention	N/A[a]	N/A[a]	4	4.6 [3.9-6.8]
				Dacarbazine	Comparator	N/A[a]	N/A[a]	4	1.6 [1.4-3.0]
TA347 (2015)	2015	PFS and OS both arms		Nintedanib + Docetaxel	Intervention	5.7	N/A[b]	12.3	6.6 [4.3-11.6]
				Docetaxel	Comparator	9.9	2.5 [2.4-3.5]	12.3	4.6 [3.9-9.8]
TA396 (2016)	2016	PFS and OS both arms[c]	R package `SiZer` used	Dabrafenib + trametinib	Intervention	18.6	18.3 [10.7-23.7]	11.8	16.8 [11.5-24.1]
				Dabrafenib or Vemurafenib	Comparator	3.2	N/A[c]	12.5	13.6 [12.9-16.1]
TA428 (2017)	2017	OS both arms[c]	Visual inspection	Pembrolizumab	Intervention	12	12.2 [10.7-12.7]	N/A[a]	N/A[a]
TA447 (2017)	2017	OS both arms[c]	Visual inspection	Pembrolizumab	Intervention	5.1	N/A[b]	N/A[a]	N/Aa

TA - Technology Appraisal; PEM - Piecewise Exponential model; OS - Overall Survival; PFS - Progression Free Survival; N/A - Not applicable.

* Mean final change-point and 95% credible interval

N/A indicates that there was no result presented because:

[a] Outcome not analysed by the B&B approach

[b] No change-point was detected for survival outcome

[c] Treatment switching occurred for comparator arm, therefore not possible to accurately reconstruct OS data.

4.3.2 Prediction accuracy of survival models

Subsequent DCOs became available after the original submission in four of the TAs using the B&B approach: TA269, TA396, TA428, and TA447. For TA269, TA396 and TA428 the updated DCO was the BRIM-3 study (Chapman et al., 2011), a pooled analysis of COMBI-v and COMBI-d trial data (Robert et al., 2019), and updated data for KEYNOTE-10 clinical trial (Herbst et al., 2021) respectively. However, the BRIM-3 study only presented an updated DCO for OS, and OS data with original cut off was not modelled using the B&B approach in TA269. As mentioned previously, the updated DCO relevant for TA447 was used in the subsequent technology appraisal TA531. Hence, we have updated DCOs for the survival outcomes in TA396, TA428, and TA447. Table 4.2 presents the (average) difference in RMST between the maximum follow-up of the updated DCO and predicted survival of the parametric models. In terms of accuracy to the updated DCO (measured by lower absolute difference in RMST) the PEM performed well relative to the other parametric models, however, for TA396 all parametric models underestimated survival, while for TA447 the updated DCO was still quite immature. Figures 4.1 and 4.2 present survival, cumulative hazard, and hazard functions for the DCOs relevant for TA428. The PEM model produces an accurate extrapolation to the updated data (Figure 4.1), and the PEM model had the lowest (best) WAIC (based on original data). The Kaplan-Meier survival function before dashed vertical line indicates earlier data-cut, while Kaplan-Meier survival function afterwards is the long-term follow up. Green diamonds refer to change-points identified by PEM. For clarity only the top three best fitting standard parametric models are presented in this figure. Because the B&B approach uses the cumulative hazard function to estimate the timepoint after which constant hazards are assumed, we present the cumulative hazard function of the earlier KEYNOTE-10 data (used in TA428) and estimates of the hazard function in Figure 4.2. Although the plot of the cumulative hazard function (Figure 4.2) is suggestive of a change in the hazards, the timepoint at which this occurs is unclear. There is a plateau in the function from $\approx$ 17 months, however, this was an artifact of a small sample size and not a true indication of a change in hazards, as confirmed by the later DCO. This was correctly identified by the PEM, with the final change-point identified before this timepoint at 12 months. The hazard function was estimated using two approaches for comparison in Figure 4.2B. The first approach uses B-splines from a generalized linear model perspective to estimate a smoothed hazard function along with 95% confidence regions (see R package `bshazard` by Rebora et al. (2018)), coloured in blue (longdash line) with confidence regions in blue. The second approach plots the posterior expectation of the hazard function from the PEM along with the 95% credible regions (in purple). Additional figures for the other TAs are provided in Appendix B.3.

Table 4.2: Goodness of fit for parametric survival models to original data and difference between predicted survival from parametric survival models to the long-term follow up data.

Original TA	Updated Data Cut-Off	Outcome	Parametric Model	WAIC	RMST	$E[\lvert\text{RMST}_{KM} - \text{RMST}_{Model}\rvert]$
TA396	COMBI Pooled	OS	Piecewise Exponential	2279.54	35.07	3.73
			Royston-Parmar 1 knot	2274.56	34.68	4.12
			Exponential	2343.35	33.83	4.97
			Log-Normal	2283.64	33.11	5.69
			Log-Logistic	2293.7	32.19	6.61
			Generalized Gamma	2292.67	31.23	7.57
			Gamma	2301.99	29.71	9.09
			Weibull	2310.25	29.27	9.53
			Gompertz	2335.74	28.69	10.11
			Follow-up Kaplan-Meier	N/A	38.85	N/A
		PFS	Royston-Parmar 1 knot	2658.45	22.96	1.67
			Piecewise Exponential	2539.88	21.97	2.6
			Gompertz	2735.91	21.32	3.25
			Log-Normal	2676.1	20.76	3.8
			Log-Logistic	2689.92	20.42	4.14
			Exponential	2736.44	19.73	4.83
			Generalized Gamma	2698.07	19.72	4.84
			Weibull	2731.17	18.82	5.73
			Gamma	2724.33	18.63	5.92
			Follow-up Kaplan-Meier	N/A	24.65	N/A
TA428	KEYNOTE-10	OS	Piecewise Exponential	1314.66	21.65	1.25
			Log-Normal	1316.25	20.47	2.36
			Log-Logistic	1317.73	19.66	3.17
			Generalized Gamma	1318.17	17.47	5.36
			Exponential	1324.93	17.29	5.54
			Royston-Parmar 1 knot	1320.92	17.26	5.57
			Gompertz	1326.44	16.83	6
			Gamma	1321.28	16.06	6.76
			Weibull	1322.71	15.78	7.05
			Follow-up Kaplan-Meier	N/A	22.99	N/A
TA447	TA531	OS	Log-Normal	387.31	21.47	0.79
			Gompertz	390.15	21.49	0.79
			Royston-Parmar 1 knot	388.78	21.51	0.79
			Log-Logistic	389.4	20.95	0.84
			Generalized Gamma	389.48	20.78	0.9
			Exponential	389.1	20.5	1.03
			Weibull	390.79	20.49	1.04
			Piecewise Exponential	389.01	20.44	1.07
			Gamma	390.68	20.4	1.09
			Follow-up Kaplan-Meier	N/A	21.32	N/A

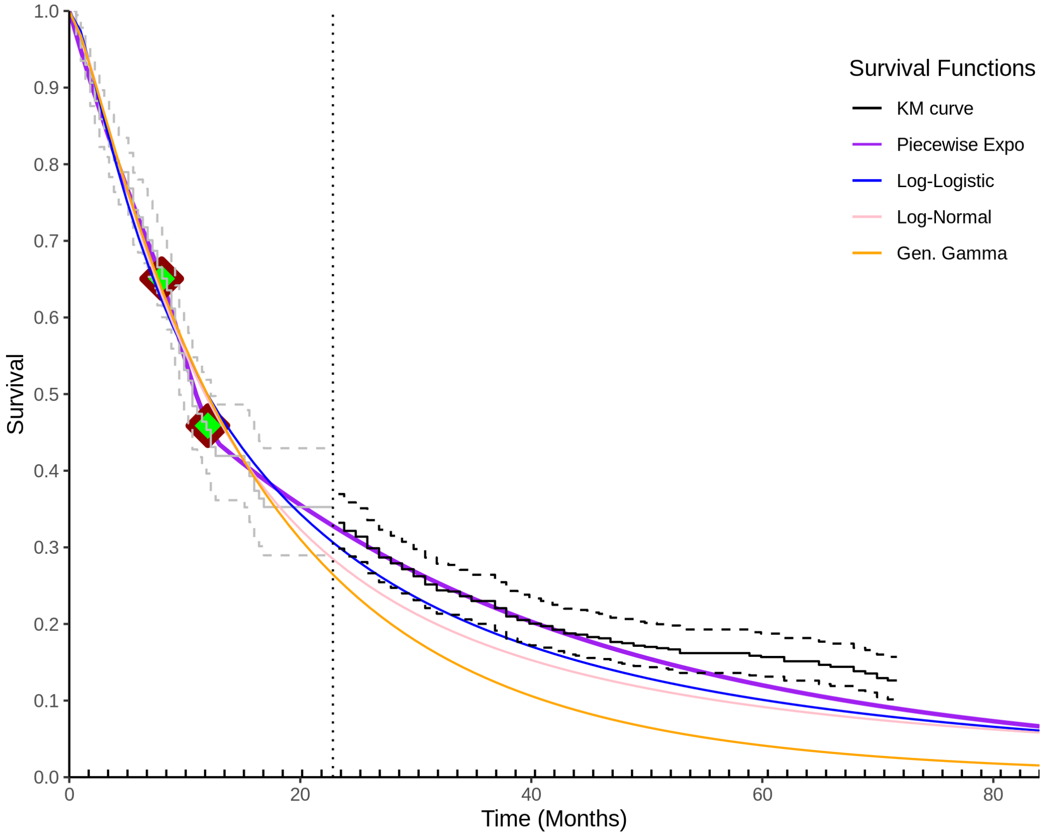


Figure 4.1: Long-term survival probabilities for various models compared to long-term data from KEYNOTE-10

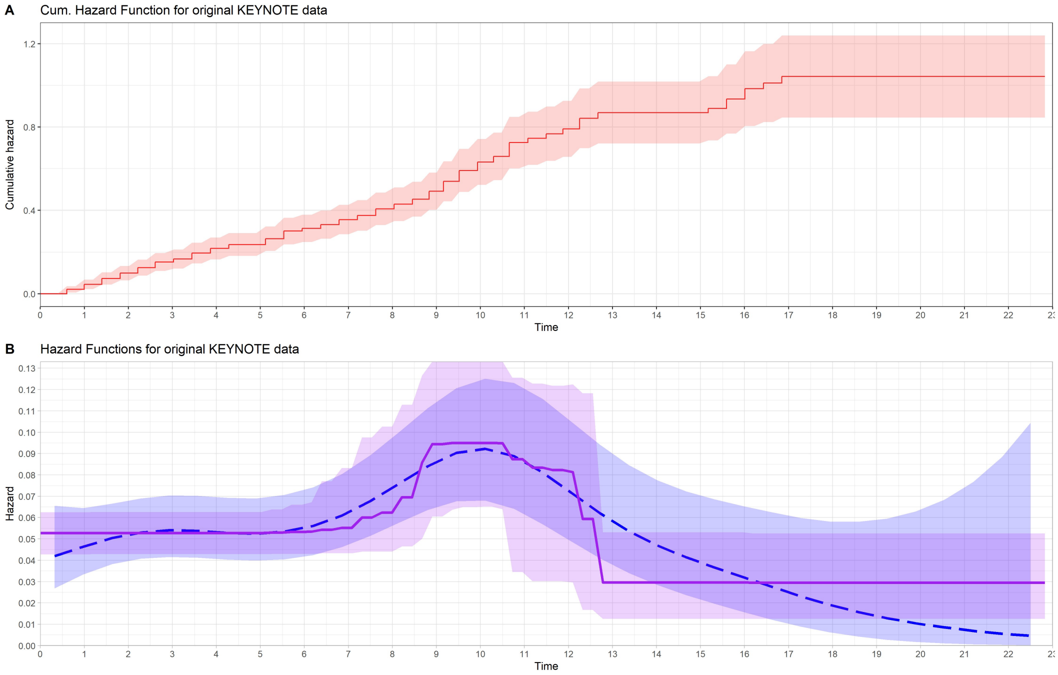


Figure 4.2: Cumulative hazard & hazard functions for original KEYNOTE-10 data

4.4 Discussion

In this chapter we reviewed the B&B approach to extrapolating survival outcomes and highlighted some of its limitations. We address each of the four identified limitations of the B&B approach with a Bayesian approach to determine the number and location of change-points in a hazard function. We also identify when there is limited evidence of a change-point in the data. The primary advantage of our method is that the specification of the change-point is data driven rather than by subjective visual inspection, a criticism raised by the ERG in TA428 (2017). By employing a Bayesian approach, uncertainty around the number of change-points is automatically computed and is described within a probabilistic framework, allowing us to assess the relative evidence of alternative change-point models. This fully parametric approach has the advantage of allowing the data to be accurately modelled while also allowing for comparison in terms of goodness of fit with other survival models described in NICE TSD 14 and 21. There are several studies which assess the performance of parametric survival models in predicting long-term survival by comparing predictions estimated from earlier DCOs (sometimes generated artificially) with later DCOs (Cooper et al., 2022; Roze et al., 2023; Bullement et al., 2019; Klijn et al., 2021; Kearns et al., 2019). Some studies have suggested that both standard and spline parametric models underestimate survival in trials of oncology immunotherapies (Cooper et al., 2022; Bullement et al., 2019). In contrast, Kearns et al. (2019) found only four of the eleven flexible parametric models considered gave an estimate of lifetime survival which was clinically plausible with the rest overestimating survival. In two of the three TAs for which later DCOs became available, we found that extrapolated survival estimated from PEMs was closest to the longer-term follow-up, relative to the other parametric models. There are, however, counter-examples in which PEMs would not be appropriate. Klijn et al. (2021) reported that the survival estimated by the B&B approach underestimated survival because the hazards continued to decrease beyond follow-up. The accuracy of survival predicted by the PEM depends on the validity of the long-term constant hazards assumption for the disease process. With some diseases/therapies, this constant hazard assumption will not be valid, however, in others it may be plausible. The PEM is a parsimonious modelling choice, assuming hazards observed in the trial are our best estimate for hazards beyond the trial. In contrast, other parametric models will typically assume that the trend (increasing/decreasing hazards) observed in the trial will continue beyond the trial, potentially leading to scenarios whereby the extrapolated hazard is markedly different to the observed one. Estimated survival from PEM (or any parametric) model should be adjusted for GPM so that the increasing hazard associated with ageing is modelled. It should be noted that this analysis has several limitations. Firstly, we identified usage of the B&B approach based on a previous review which had a timeframe from 2011-2017 and whose scope was restricted to NICE TAs. Additionally, we reconstructed patient level data from KM plots which inevitably is less accurate than analysing the original data. Both NICE TSD 14 and 21 briefly discuss piecewise models, highlighting limitations around the arbitrary number and location of change-points. In the context of survival extrapolation, this work is the first consider to robust estimation of change-points in a piecewise exponential model, additionally allowing comparison with other parametric survival models. The framework presented in this chapter could be extended to estimate other parametric (e.g., Weibull) change-point models in Chapter 5 along with the inclusion of covariates to jointly model the intervention and comparable.

4.5 Conclusions

In this chapter we applied the change-point model developed in Chapter 3 to a number of previously conducted Technology Appraisals which used the B&B approach. If disease specific hazards can be assumed to be relatively constant, or if there is no prior understanding of the trend of these hazards then the PEM may be considered for survival extrapolation. Although unlikely that the disease specific hazards will remain strictly constant over the course of the extrapolated horizon, it is often not known whether these hazards will increase or decrease as is assumed by other parametric models. While the hazard from the disease process is assumed to be constant in the PEM, survival extrapolations generated from PEM should include GPM so that the marginal/total hazard is increasing over time. We encourage practitioners who are using the B&B approach to model survival for HTA to instead employ the approach presented in this study.

While constant hazard extrapolations may be suitable in certain scenarios, there exist numerous situations where they are not considered plausible. The inability of the collapsing sampler to jointly model the treatment and comparator is a clear limitation of the approach developed in Chapter 3 and applied to real world situations in this chapter. In Chapter 5 we estimate more complex change-point models (i.e. Weibull survival models which can accommodate covariates) and show how these models can consider a variety of scenarios relating to changes in relative treatment effects.

5 More Complex Change-point Survival Models

5.1 Introduction

In Chapter 3 we have seen the application of a specific class of change-point models which assume a constant hazard within each segment. In Chapter 4 we described applications of this change-point model to a variety of technology appraisals and how the piecewise exponential change-point model is a more appropriate extrapolation strategy than manually selecting a point after which constant hazards are extrapolated (i.e. the Bagust & Beale approach).

Estimating the piecewise exponential model using a collapsed RJMCMC scheme is computationally efficient and because the marginal likelihood of the exponential model has an analytical form there is good mixing between models of differing dimensions (i.e. models with different change-point locations). However, the relative probabilities of the change-point models estimated by RJMCMC are proportional to the marginal likelihood of the data, a quantity which is sensitive to the priors placed on the parameters. The piecewise exponential model described in Chapter 3 mitigated this sensitivity by allowing the β hyperparameter to be informed by data and allows for a more robust estimation of the change-point model probabilities with respect to differences in the prior.

There are however, a number of limitations of the collapsed PEM model. Firstly, the constant hazard assumption can be considered restrictive, and we may want to consider models which allow for decreasing/increasing hazards within each segment. Secondly the collapsed RJMCMC scheme requires the evaluation of the marginal likelihood which is only available in an analytical form for an exponential likelihood with no covariates. The absence of covariates does not allow us to jointly model the intervention and the comparator, an important consideration in HTA. Relatedly the lack of covariates means general population mortality cannot be modelled using the internally additive hazards approach, an approach which is recommended by Rutherford et al. (2020). Another limitation for the collapsed change-point models is that the change-points are restricted to a finite number of times (in our implementation event times). Although the likelihood of the change-points are maximized at event times, there is no clinical reason why change-points are constrained to be at event times.

One approach to estimating models with covariates would be to design a bijective function to allow movement between the models of different dimensions (Chapple et al., 2020). In practice it can be difficult to design such a function which will allow for reasonable mixing between the models of differing dimensions. Because the survival data available at HTA assessment are typically immature it is unlikely that a large number of change-points should be required, particularly as the final interval would then contain fewer events to estimate the model which will be used for extrapolation purposes. In situations where the number of candidate models is small (i.e. with respect to the number of change-points) it is more efficient to estimate each of the models and make a selection based on goodness of fit. To obtain a result close to that which would be obtained by RJMCMC, one could evaluate the marginal likelihood, however, goodness of fit statistics (which are typically more easily evaluated) could also allow for model averaging across the candidate models (Jackson et al., 2010).

It's worth motivating the rationale for estimating change-point models for decision modelling in HTA. For the scenarios described in the previous chapter the purpose was to identify a time-point (statistically) after which a constant hazard can be assumed appropriate. While the constant hazard assumption will typically not hold over the course of the extrapolated time horizon, it can be an appropriate assumption if we believe the hazards during the trial are a reasonable estimate of the long-term hazards and lack information on the potential trend of the long-term hazard function over time. Considering a Weibull change-point model rather than a piecewise exponential model could allow us to model the data with fewer change-points and we could still constrain the final interval to contain a constant hazard (as the exponential is a Weibull model with the shape parameter set to 1).

Previous authors have described highly flexible parametric models which allow for the modelling of complex hazard functions, however even these models typically do not accommodate specific scenarios which health economic modellers might wish to consider. Some such examples are detailed below and relate to the joint modelling of the treatment and intervention:

- Treatment Delay (TD) - Hazard function for both treatment and comparator is the same until a certain timepoint
- Loss of Treatment Effect (LTE) - Hazard function for treatment and comparator is the same after a certain timepoint

- Converging Treatment Effect (CTE) - Hazard ratio converges over time to one (i.e. equal hazards)

For certain treatments, it has been hypothesized that a delay may be observed after the initiation of treatment before differences in the survival times of the groups become apparent. In the opposite scenario, after a period of initial benefit, the treatment effect may no longer be observed and the hazard function for both treatment and comparator are equal. A related scenario is the potential for a measure of treatment effect such as the hazard ratio (HR) to change after a timepoint and converge to 1 in a smooth fashion. Various applications of CTE and LTE in NICE Technology Appraisals are documented by Kamgar et al. (2022), however, the timepoint after which the change in treatment effect occurs is uncertain often arbitrary chosen.

As we demonstrate in this chapter, each of these scenarios relating to changes in treatment effects can be modelled with change-point models. As in the Chapters 3 and 4 we will estimate these models assuming that the change-point locations are unknown. In Section 5.2 we will describe the notation of the generic change-point survival model and the three use cases we have highlighted above. In Section 5.5 we describe three datasets to which we apply specific change-point models whose results are presented in Section 5.6.

5.2 General Specification of a Parametric Change-point Survival Model

5.2.1 Likelihood of Parametric Change-point Survival Models

Distinct from the change-point model introduced in Chapter 3 which assumed change-points at event times (i.e. discrete timepoints), in this section we consider continuous change-point models. As in Chapter 3 we define $\mathbf{t}_{1:n}$ as a vector of n time ordered survival times. Multiple change-points can be denoted as a vector $\boldsymbol{\tau}_{1:k}$ (and individual change-points denoted with a single subscript e.g. τ_j), with these k change-points partitioning time into $k+1$ segments. In order partition time into the $k+1$ segments we also require boundary change-points τ_0 and τ_{k+1}. We define $\tau_0 = 0$ and $\tau_{k+1} \geq t_n$, however, as these quantities are not parameters to be estimated by the model we will not include them in the model notation.

Owing to the potential for covariates, we require that each individual and interval has a specific hazard function $h(t_i, \boldsymbol{\theta}_{ij})$. For each individual we require the cumulative hazard function up until time t_i. This includes the cumulative hazard for the interval between t_i

(assuming it occurs in the jth interval) to the previous change-point τ_{j-1} denoted as $H((t_i - \tau_{j-1}), \theta_{ij})$. The cumulative hazard function for any previous intervals is also required and is denoted as $H(\tau_g - \tau_{g-1}, \boldsymbol{\theta}_{ig})$. The likelihood of the change-point model can be formulated as follows

$$L(\boldsymbol{\tau}_{1:k}, \boldsymbol{\theta}|\mathbf{t}_{1:n}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k+1} h(t_i, \boldsymbol{\theta}_{ij})^{\delta_{ij} v_i} \exp\left\{ -\delta_{ij} \left[H((t_i - \tau_{j-1}), \boldsymbol{\theta}_{ij}) + \sum_{g=1}^{j-1} H(\tau_g - \tau_{g-1}, \boldsymbol{\theta}_{ig}) \right] \right\} \right\}. \tag{5.1}$$

with $v_i = 1$ if the ith subject was observed to have an event and 0 otherwise (censored). If the ith subject's time (either censored or an event) is within the jth interval (mathematically $t_i \in (\tau_{j-1}, \tau_j]$), then $\delta_{ij} = 1$ and 0 otherwise.

Furthermore $\boldsymbol{\theta}_{ij}$ is a vector of parameters for each individual which in the case of the Weibull model are the shape and scale parameters which can have covariates placed upon them. For example, let $\boldsymbol{\theta}_{ij} = \{m_{ij}, a_{ij}\}$ with m, a scale and shape parameters respectively. To model covariates, we introduce matrices $\boldsymbol{\beta}_m = [\boldsymbol{\beta}_{m_1} \dots \boldsymbol{\beta}_{m_{k+1}}]$ and $\boldsymbol{\beta}_a = [\boldsymbol{\beta}_{a_1} \dots \boldsymbol{\beta}_{a_{k+1}}]$ whose columns are a $p \times 1$ vector each representing the coefficients of one of the $k+1$ intervals. For the jth interval $m_{ij} = \exp(\mathbf{Z}_{ij}\boldsymbol{\beta}_{m_j})$ in which the scale (location) parameter depends on a vector of covariate values $\mathbf{Z}_{ij}$ of size $1 \times p$ and and the coefficients of $\boldsymbol{\beta}_{m_j}$. The individual shape parameter is calculated as $a_{ij} = \exp(\mathbf{Z}_{ij}\boldsymbol{\beta}_{a_j})$.

If we do not specify a treatment effect for the shape parameter we obtain a proportional hazards model, and for each of the p covariates (naturally excluding the intercept) the interval specific hazard ratio for the qth covariate is $\mathrm{HR_{jq}} = \exp(\beta_{\mathrm{jq}})$.

5.2.2 Data format for Parametric Change-point Survival Models

To help clarify the notation in Section 5.2.1 we will provide illustrative examples of various change-point scenarios with one change-point i.e. $k = 1$, which is typically sufficient to fit the observed data. For the purposes of illustration we will provide a dataset with 5 observations, three assigned to treatment and two assigned to a comparator (or baseline) along with ages of each patient (see Code Chunk 1). As per the notation from the previous section, the individual survival time is t_i and censoring indicator is v_i. In this dataset we have two covariates which we will include in the model, treatment status and age. The age variable is transformed (or scaled) to have a mean of zero and a standard deviation of one. Scaling variables can improve MCMC sampling efficiency, and it also simplifies prediction for the population at the mean age, as the coefficient for age can be omitted. Furthermore, it can be more straightforward to define priors for the coefficients, as they represent the change in the log hazard ratio for a change of one standard deviation from the mean. While redundant for this dataset, we include an integer variable indicating the patient's ID which is necessary for the modified dataset introduced later.

The ID variable is defined in the dataset (Code Chunk 1) as `id`. Individual survival time is defined as `time` while the censoring indicator is defined as `status`. Treatment status is defined by the `trt` variable with `trt` $= 1$ indicating treatment and `trt` $= 0$ the comparator. The age variable is defined as `age` and the corresponding scaled variable is `age_scale`.

```
id time status trt   age age_scale
1  0.08      1   1 77.64      1.14
2  0.14      0   0 67.75     -0.10
3  1.44      0   1 73.45      0.61
4  2.11      1   1 56.36     -1.52
5  3.32      0   0 67.52     -0.13
```

Listing 1: Example Simulated Dataset

For the dataset given in Code Chunk 1 we assume a single change-point at timepoint equal to 1 (i.e. $\tau_1 = 1$). We require a dataset that provides information about the time an individual spends in each interval. This includes information on the complete intervals $(\tau_{g-1}, \tau_g]$ and the final interval $(\tau_{j-1}, t_i]$ which is required for the likelihood in Equation 5.1. Furthermore, we require the values of the covariates for each individual for a specific interval, i.e. the covariate vector $\mathbf{Z}_{ij}$. It should be noted that in this example these covariate values will not change for an individual across different intervals, however, their associated coefficients $\boldsymbol{\beta}_{m_j}, \boldsymbol{\beta}_{a_j}$ can potentially change.

In order to include this information we use a data format known as the "counting process" format in which each row refers to a specific interval for an individual. The individual will have a number of intervals equal to the number of change-points for which the individual's observed time is greater than the change-point, i.e. $\sum_{j=0}^{k+1} s(t_i - \tau_j - c)$ where $s(x)$ is the unit step function which returns a value of 1 if $x \geq 0$ and 0 otherwise. The unit step function requires $-c$, where c is a very small number as the individual's survival time should be strictly greater than the change-point τ_j in order to include the jth $+ 1$ interval.

For each complete interval, i.e. $t_i > \tau_j$ where $j > 0$, the dataset will contain the start time and end time for the interval which is τ_{j-1} and τ_j respectively. For the final interval the end time is simply the individual's survival time t_i. This final interval has $\delta_{ij} = 1$ with all previous intervals equal to zero. The covariate vector $\mathbf{Z}_{ij}$ is also included for each interval.

The advantage of this dataset format is that the calculation of the likelihood for the change-point is more straightforward. Given the values of the coefficients $\boldsymbol{\beta}_{m_j}, \boldsymbol{\beta}_{a_j}$ and covariate vector $\mathbf{Z}_{ij}$ we can calculate the individual interval specific values of the shape and scale parameters (i.e. a_{ij}, m_{ij}). We can then calculate the cumulative hazard for each interval, while for the individual's final interval, we additionally calculate the hazard function. With these quantities we can then calculate the likelihood contribution for each individual based on their censoring status.

The survival data presented in the "counting process" format are presented in Code Chunk 2.The column that indicates which subrecords belong to which specific patient is defined as `id` in the dataset, while the interval to which the record is associated is denoted as `Interval`. The start and end times of the interval are denoted as `tstart` and `tstop` respectively. The column indicating the event/censoring status is the product of the censoring indicator from the original dataset and the indicator for an individual's final interval i.e. $v_i \times \delta_{ij}$. By definition this will be zero for all intervals before an individual's final interval, as clearly an individual must have survived at least until their final interval. This variable is denoted as `status` in the dataset. In order to define the covariate vector $\mathbf{Z}_{ij}$ we require an intercept which in the dataset is defined as `Intercept` and is simply a column of ones. The other two components of $\mathbf{Z}_{ij}$ are the treatment status and age scaled which are defined as `trt` and `age_scale` (as in Code Chunk 1).

For example, the first individual (denoted with `id` $= 1$) has $t_i < 1$, therefore, only has one record with `tstart` and `tstop` as 0 and t_i respectively. The third individual has $t_i > 1$ and has two subrecords. The first subrecord corresponding to the first (complete) interval has `tstart` $= 0$ and `tstop` $= 1$, while the second and final interval has `tstart` $= 1$ and `tstop` $= t_i$.

```
 id Interval tstart tstop status Intercept trt age_scale
  1        1      0  0.08      1         1   1      1.14
  2        1      0  0.14      0         1   0     -0.10
  3        1      0  1.00      0         1   1      0.61
  3        2      1  1.44      0         1   1      0.61
  4        1      0  1.00      0         1   1     -1.52
  4        2      1  2.11      1         1   1     -1.52
  5        1      0  1.00      0         1   0     -0.13
  5        2      1  3.32      0         1   0     -0.13
```

Listing 2: Example Dataset in a counting process format

```
beta_scale
          Interval-1 Interval-2
Intercept       -0.5       -0.5
trt             -0.2        0.0
age_scale        0.1        0.1

beta_shape
          Interval-1 Interval-2
Intercept       -0.4       -0.4
trt              0.0        0.0
age_scale        0.0        0.0

 tstart tstop status id Interval Intercept trt age_scale scale shape
      0 0.08      1  1        1         1   1      1.14  0.56  0.67
      0 0.14      0  2        1         1   0     -0.10  0.60  0.67
      0 1.00      0  3        1         1   1      0.61  0.53  0.67
      1 1.44      0  3        2         1   1      0.61  0.64  0.67
      0 1.00      0  4        1         1   1     -1.52  0.43  0.67
      1 2.11      1  4        2         1   1     -1.52  0.52  0.67
      0 1.00      0  5        1         1   0     -0.13  0.60  0.67
      1 3.32      0  5        2         1   0     -0.13  0.60  0.67
```

Listing 3: $\boldsymbol{\beta}$ covariate matrices for shape and scale parameters and updated dataset

5.2.3 Change-point scenarios based on covariate values

By restricting various covariates (i.e. elements of $\boldsymbol{\beta}_{m_j}, \boldsymbol{\beta}_{a_j}$) to be equal to 0 or equal across the intervals we can specify many different change-point models.

In Code Chunk 3 we have specified covariates for the scale parameter based on treatment status and age. The effect of age is constant across the intervals while the treatment effect varies across intervals, in fact because it is zero in the second interval the hazards are equal to the comparator i.e. LTE model. The opposite scenario is where we constrain the treatment coefficient to be zero in the first interval yielding a TD model.

The shape parameter is constant across the intervals (and more generally not subject to a treatment effect) resulting in a proportional hazards change-point model. Because the intercept for the scale parameter is constant across intervals, the baseline hazard is continuous at the change-point, however, if the intercept is allowed to vary across intervals, it is non-continuous but still a proportional hazard model.

In Figure 5.1 we present four possible scenarios to jointly model the hazard function for a change-point model in which the HR for the treatment was <1 up until (and including) the change-point and greater than 1 after the change-point. Figure 5.1-A illustrates the scenario previously described in which only the HR of the treatment changes after the change-point and so the baseline (comparator) hazard function is continuous. In Figure 5.1-B the intercept also changes across intervals so that there is a different baseline hazard function for each interval. Figure 5.1-C extends 5.1-B so that the shape parameter also changes across intervals. It is worth highlighting that scenarios A-C still assume proportional hazards while the final scenario in Figure 5.1-D assumes different shapes for both treatment arms after the change-point and therefore is no longer a PH model for the interval after the change-point.

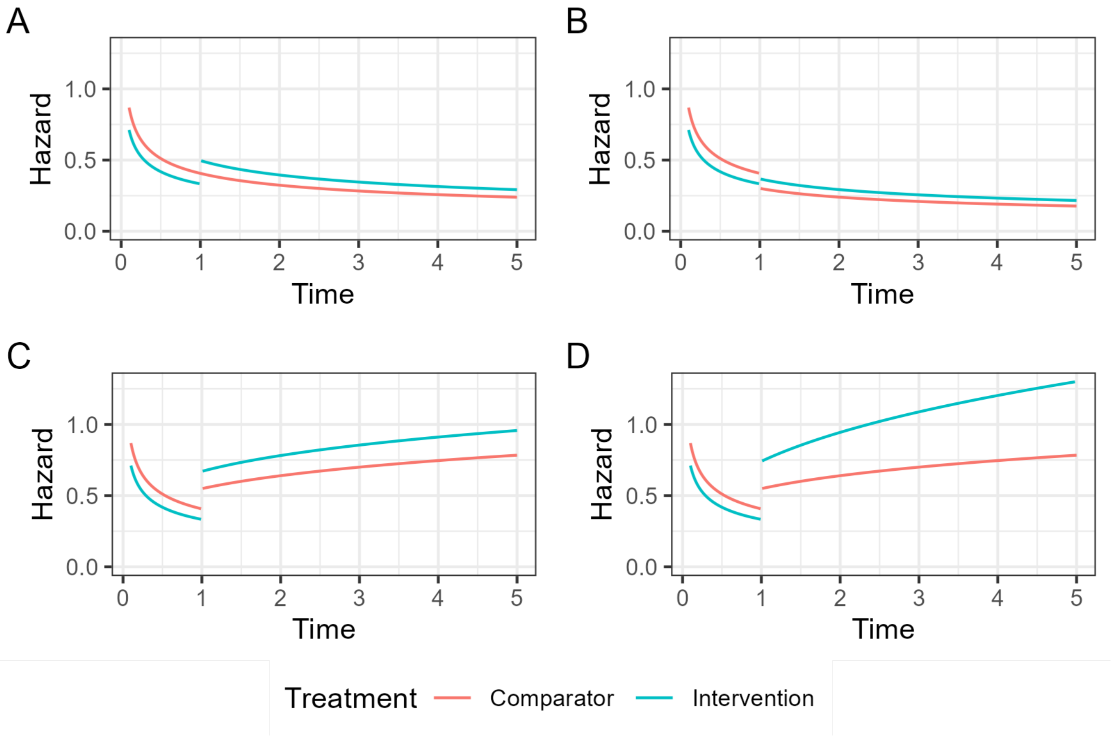


Figure 5.1: Various scenarios for modelling the hazard function with a change-point

5.2.4 Change-point scenarios with Convergence of the Hazard Ratio

In the previous scenario we have considered a step change in the HR, however, we may also allow the HR of the treatment arm to converge in a continuous manner to the comparator or baseline hazard i.e. CTE models. For a converging hazards model we consider a change-point τ_{wane} after which the hazard ratio for the treatment from the previous interval ($\text{HR}_{\text{initial}}$) begins to wane (i.e. converge to 1 over time). The HR for any time after τ_{wane} is

$$\text{HR}(t) = 1 - (1 - \text{HR}_{\text{initial}}) \exp(-\omega(t - \tau_{\text{wane}})), \tag{5.2}$$

were ω is a constant rate at which the HR converges to 1. Figure 5.2 shows various HRs with different values of ω.

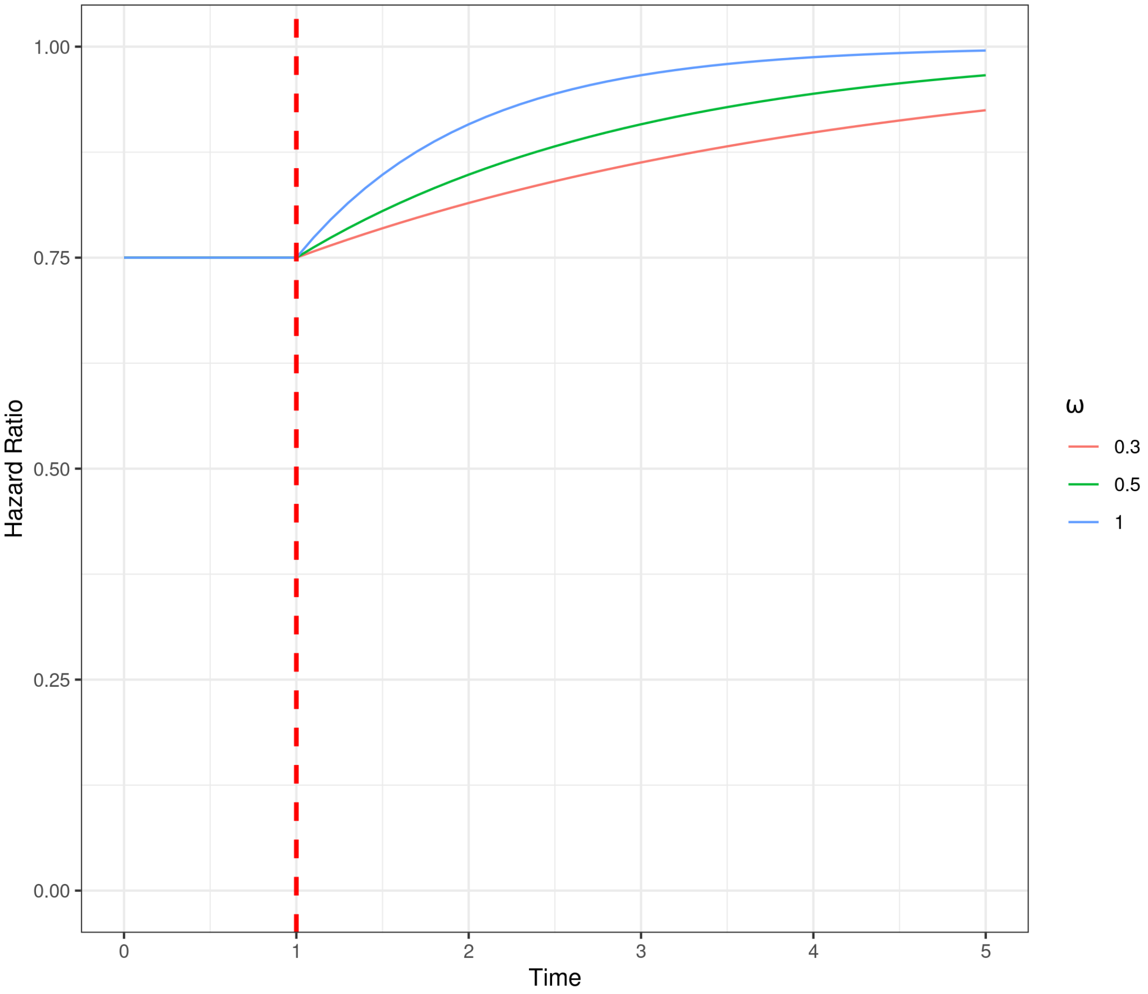


Figure 5.2: Hazard Ratios for different rates of convergence ω; $\tau_{\text{wane}} = 1, \text{HR}_{\text{initial}} = 0.75$

In order to estimate the cumulative hazard function for the treatment arm, we need to evaluate $\int_{\tau_{\text{wane}}}^{t} \text{HR(t)h}_{\text{baseline}}(t)dt$ with $\text{h}_{\text{baseline}}(t)$ being the baseline hazard. For the Weibull model the indefinite integral is $am_{\text{baseline}}(t^a)\Big(1/a - (\Gamma(a,\omega t)(\text{HR}_{\text{initial}} - 1)\exp(\omega\tau_{\text{wane}}))/(\omega t)^a\Big) + C$ for the interval beyond τ_{wane}. m_{baseline} refers to the baseline scale parameter with the shape a common for both intervention and comparator arm and $\Gamma(a,\omega t)$ is a the upper incomplete gamma function. For the exponential likelihood the integral simplifies considerably with $a = 1$.

We define the β_m matrix as before see Code Chunk 4, however, the HR for the treatment effect beyond the change-point is a function of time since the change-point, the HR for the treatment before the change-point and ω, the rate of convergence (Equation 5.2).

```
beta_scale
          Interval-1 Interval-2
Intercept       -0.5       -0.5
trt             -0.2        log(HR(t))
age_scale        0.1        0.1
```

Listing 4: β_m covariate matrix for CTE model

Each of the scenarios presented in Figure 5.3 correspond to those (proportional hazard models) presented in Figure 5.1.

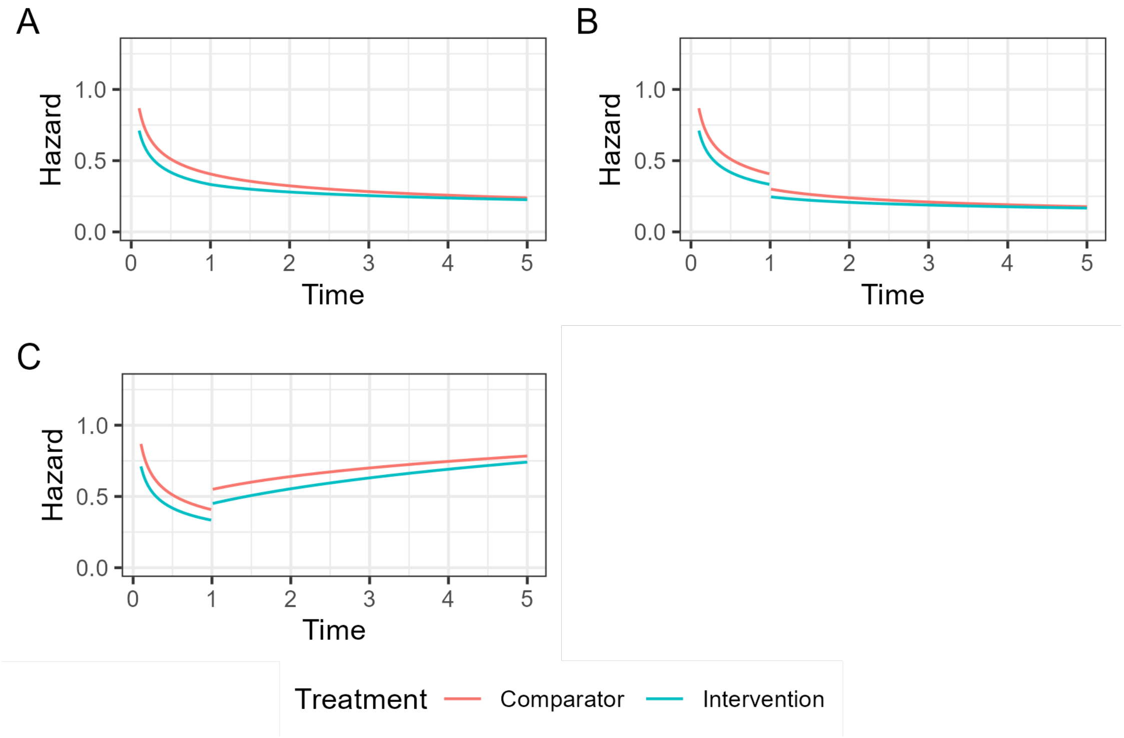


Figure 5.3: Various scenarios for converging hazard with a change-point

5.2.5 Summary

In subsequent examples we will consider the Weibull and exponential (setting the shape parameter to 1) change-point models although for the treatment delay and loss of treatment effect we could consider accelerated time factor models such as log-normal and gamma models. For the converging hazards approach it is appropriate to restrict our attention to proportional hazards models as time acceleration factors are not typically considered in applications of treatment waning. The framework we present could easily be extended to include the Gompertz model (which is also a proportional hazards model).

5.3 Estimation of Parametric Change-point Survival models

All of the models described in the subsequent sections are estimated using the JAGS statistical software program Plummer (2003). The model code to define the change-point model likelihood makes extensive use of the unit step function $s(x)$ defined in Section 5.2.2. This function is defined as the `step` function in JAGS. Using the notation defined previously we can use two unit step functions to calculate the interval which t_i is in. To test if t_i is within the jth interval we can use $s(t_i - \tau_{j-1} - c) \times s(\tau_j - t_i)$ which will return 1 if and only if t_i is in the jth interval. As in Section 5.2.2, the first unit step function requires the small constant c as the individual's survival time should be strictly greater than the lower change-point value (τ_{j-1}) to be in the jth interval. In practical terms this variable can be excluded when dealing with continuous change-points as the probability $\tau = t_i$ is 0.

Unlike custom-written MCMC samplers, JAGS has a limited number of probability distributions that can be used when defining a generative model (although sufficient for a wide range of statistical models). Additional distributions can be included, but they require knowledge of C++ to implement these extensions (Wabersich and Vandekerckhove, 2014). Because JAGS does not include distributions corresponding to change-point survival models, we need an approach to include the likelihood contribution of the data without resorting to additional programming. One approach is to use a distribution that is already available in JAGS to include the contribution of the likelihood for the change-point model, i.e., $L(\boldsymbol{\tau}_{1:k}, \boldsymbol{\theta}|\mathbf{t}_{1:n})$. Known as the "zeros trick", we create an observation $z = 0$, which is assumed to be drawn from a Poisson(ξ) distribution. Because the observation $z = 0$, the likelihood contribution is $\exp(-\xi)$. Therefore, setting $\xi = -\log(L(\boldsymbol{\tau}_{1:k}, \boldsymbol{\theta}|\mathbf{t}_{1:n}))$ produces the correct likelihood contribution for the change-point model. Note that ξ needs to be positive as it is a Poisson mean, therefore, we add a suitably large (but otherwise arbitrary) constant to ensure that it is positive.

The choice between alternative change-point (and standard parametric) models can be guided by a goodness of fit measure such as WAIC, however, consideration must also be given to the plausibility of the hypothesis underlying the statistical models.

Markov chain sampling of the parameters is achieved using slice sampling (Neal, 2003). We expect Hamiltonian MCMC (as used in Stan) would fail as the likelihood is not a smooth function of $\boldsymbol{\tau}_{1:k}$ and evaluation of the gradient required for the exploration of the posterior would not be possible. We place a vague prior for the $\boldsymbol{\beta}$ covariates for the shape and scale parameters i.e. $\mathcal{N}(\mu = 0, \sigma = 5)$.

For the change-point we assume that $\tau_{1:k}$ are even ordered statistics of $2k$ split points drawn from a Uniform distribution on $(0, \tau_{\text{max}})$ and we assume that τ_{max} is the maximum observed time. Furthermore, we define $\tau_0 = 0$ and $\tau_{\text{max}} = \tau_{k+1}$. Formally this prior distribution is $p(\boldsymbol{\tau}_{1:k}|k) = \frac{(2k+1)!\prod_{k=1}^{k+1}(\tau_k - \tau_{k-1})}{\tau_{\text{max}}^{2k+1}}$ and has been used extensively in the estimation of continuous change-point models (such as Chapple et al. (2020)) and is the continuous analogue of the discrete change-point prior presented in Chapter 3.

Because the discrete change-point prior calculates the prior probability in terms of events while the continuous prior is with respect to time there are differences in the priors. Assuming a single change-point, the density is maximized for the continuous prior at the mid-point between $(0, \tau_{\text{max}})$, while for the discrete prior the probability is maximized at the median event time. The discrete prior is dependent on the number of events and the distribution of these events. For the purposes of illustration we compare the discrete change-point prior for 100 observations simulated from an exponential distribution with $\lambda = 0.4$ and all observations > 5 censored with a continuous prior from $(0, \tau_{\text{max}} = 5)$. In contrast to the continuous prior, the discrete prior provides an asymmetrical prior with respect to time with less probability towards the maximum observable time (Figure 5.4). Because the discrete prior is based on the number and distribution of event times, this prior will be different for each dataset and dependent on the underlying distribution of event times. This could be considered a disadvantage, however, the discrete prior does have the advantage assuming a lower probability of a change-point in intervals which have few event times. For example, the continuous prior in Figure 5.4 assumes that a change-point is equally likely in the region $(0, 1]$ as $[4, 5)$ even though the later interval has much fewer event times. If a particular prior is considered more appropriate, such as a truncated normal distribution it is straightforward to amend the JAGS code to do this.

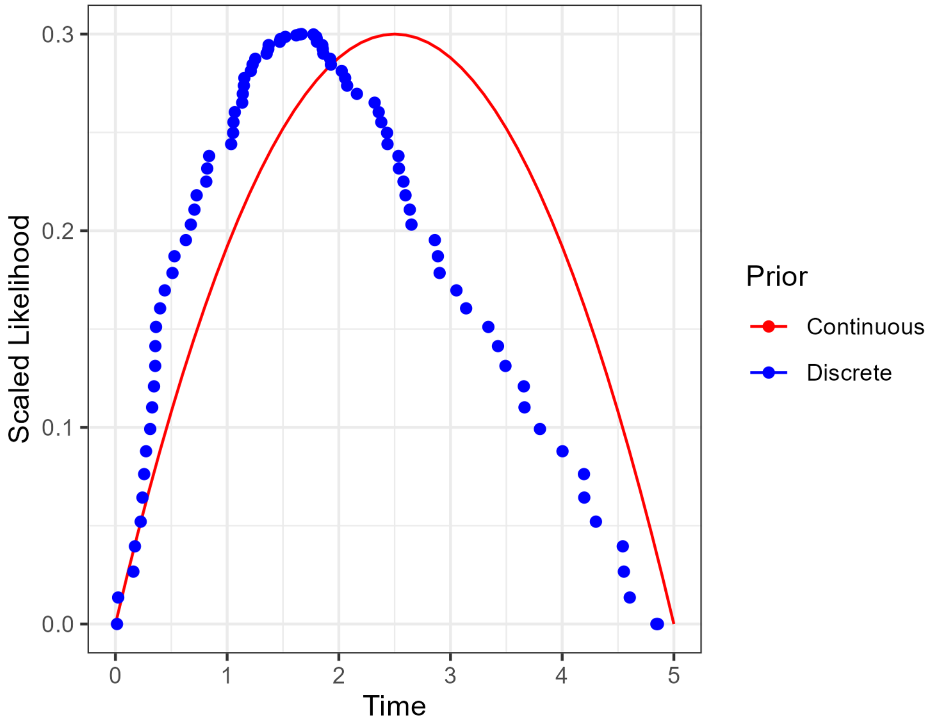


Figure 5.4: Comparison of discrete and continuous change-point priors

For each of the models whose results are presented in Section 5.6 we ran 2 chains for 55,000 iterations with an initial burnin of 5,000 and a thinning factor of 5 (thus giving us a total sample of 20,000 iterations). Convergence diagnostics were assessed using the `ggmcmc` package by Fernández-i-Marín (2016).

Codes to reproduce the analysis presented in this chapter are available on `Github` with skeleton pseudo-code described in the Appendix (Code Chunk 8) for the special case of a one change-point Weibull model with no covariates, showing both the application of the unit step function and the "zeros trick".

5.4 Simulation Study

We consider two models for generating data arising for change-point models under a variety of different parameters and sample sizes. In all scenarios we assume that the baseline scale parameter for the Weibull model before the change-point is 0.3. The model assumes common shape parameter for both timepoints being either 0.7 or 1.2 (assuming monotonically increasing or decreasing hazards). We assume the HR between the treatment and baseline is 0.25, 0.5 or 0.75. The sample size considered for each arm was assumed to be 200, 500 or 1000 (n_{samp}) and the data-cut off was 4 years after which all observations were assumed censored, with no censoring before the data-cut off.

In the first set of scenarios we assume a treatment delay. Before a change-point which is assumed to occur at 1 year, the hazards for both treatment and baseline are equal. After the change-point we assume a PH model with various HRs and baseline hazard functions investigated. The scenarios in which a monotonically decreasing and increasing baseline hazard are assumed are considered in Figures 5.5.

In the second set of scenarios we assume a loss of treatment effect after the change-point which occurs at 2 years, while before the change-point the data arise from a PH model. The scenarios in which a monotonically decreasing and increasing baseline hazard are assumed are considered in Figures 5.6.

For each combination of parameters from each scenario we simulate 100 datasets and fit the associated Weibull change-point model along with a range of standard parametric models (i.e. exponential, Weibull, gamma, Gompertz, log-logistic, log-normal, generalized gamma and Royston-Parmar model).

In Tables 5.1 we present the difference in restricted mean survival time (RMST) up to 15 years between the treatment and baseline arms divided by the difference in RMST for the true survival functions ($\text{RMST}_{\text{diff}}$), averaged over the 100 simulations. Values closer to zero indicate closer fit to the true generating process. For clarity we present results for the change-point and other parametric models with the three lowest average values of $\text{RMST}_{\text{diff}}$ across the set of scenarios.

Table 5.1: Simulation Study results - Values of $\text{RMST}_{\text{diff}}$ for Treatment Effect Delay scenarios

n_{samp}	shape	Initial HR	Change-point	Gamma	Royston-Parmar
200	1.2	0.25	0.05	0.16	0.12
500	1.2	0.25	0.05	0.15	0.11
1000	1.2	0.25	0.06	0.16	0.12
200	0.7	0.25	0.05	0.20	0.30
500	0.7	0.25	0.06	0.21	0.30
1000	0.7	0.25	0.06	0.20	0.30
200	1.2	0.5	0.06	0.10	0.07
500	1.2	0.5	0.08	0.10	0.08
1000	1.2	0.5	0.07	0.10	0.08
200	0.7	0.5	0.09	0.18	0.26
500	0.7	0.5	0.09	0.19	0.26
1000	0.7	0.5	0.08	0.19	0.27
200	1.2	0.75	0.12	0.09	0.09
500	1.2	0.75	0.12	0.09	0.08
1000	1.2	0.75	0.12	0.09	0.08
200	0.7	0.75	0.17	0.18	0.24
500	0.7	0.75	0.17	0.19	0.25
1000	0.7	0.75	0.17	0.20	0.25

A number of points are worth noting. For each of the treatment delay scenarios we

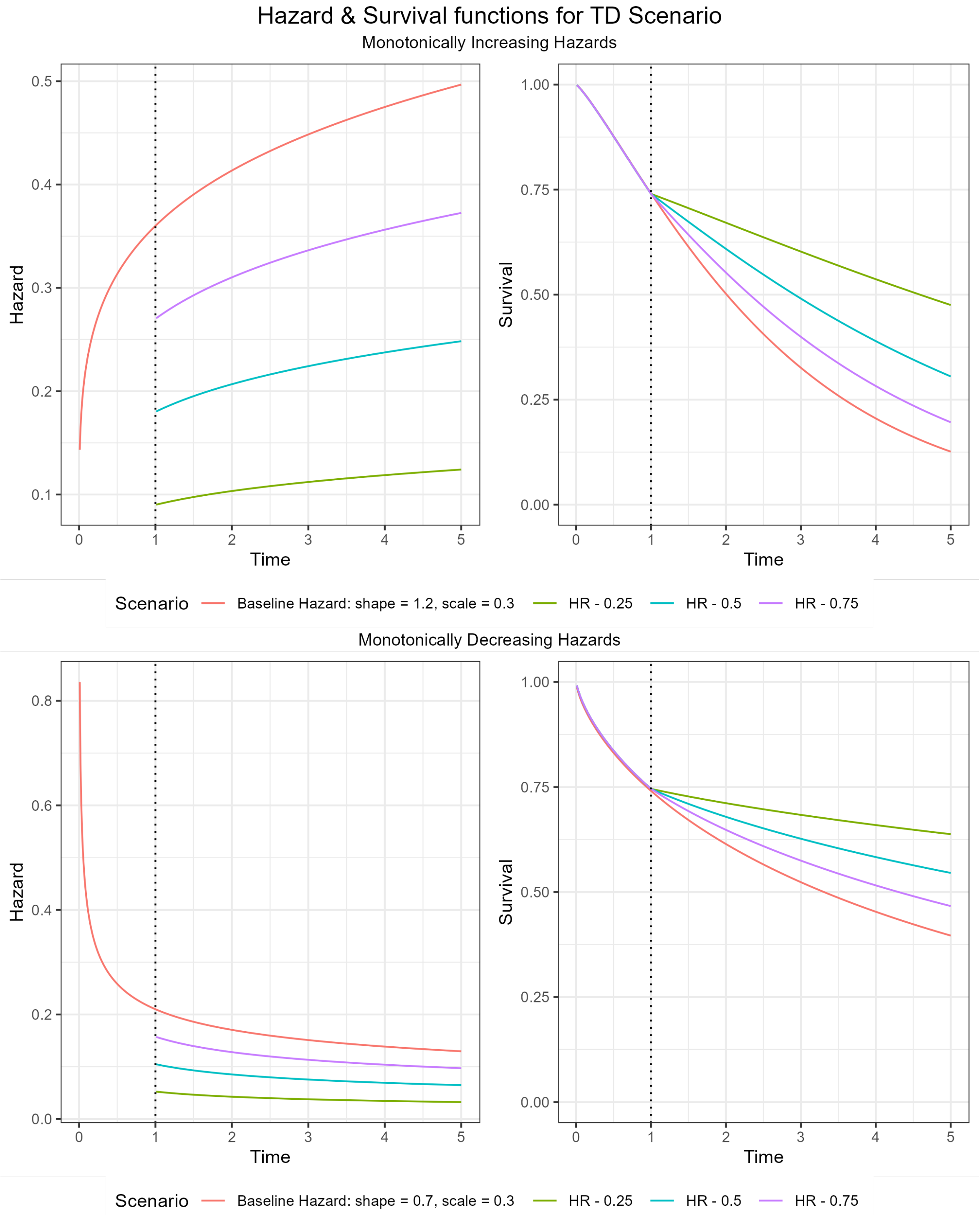


Figure 5.5: Hazard and Survival functions for Treatment Delay Scenarios

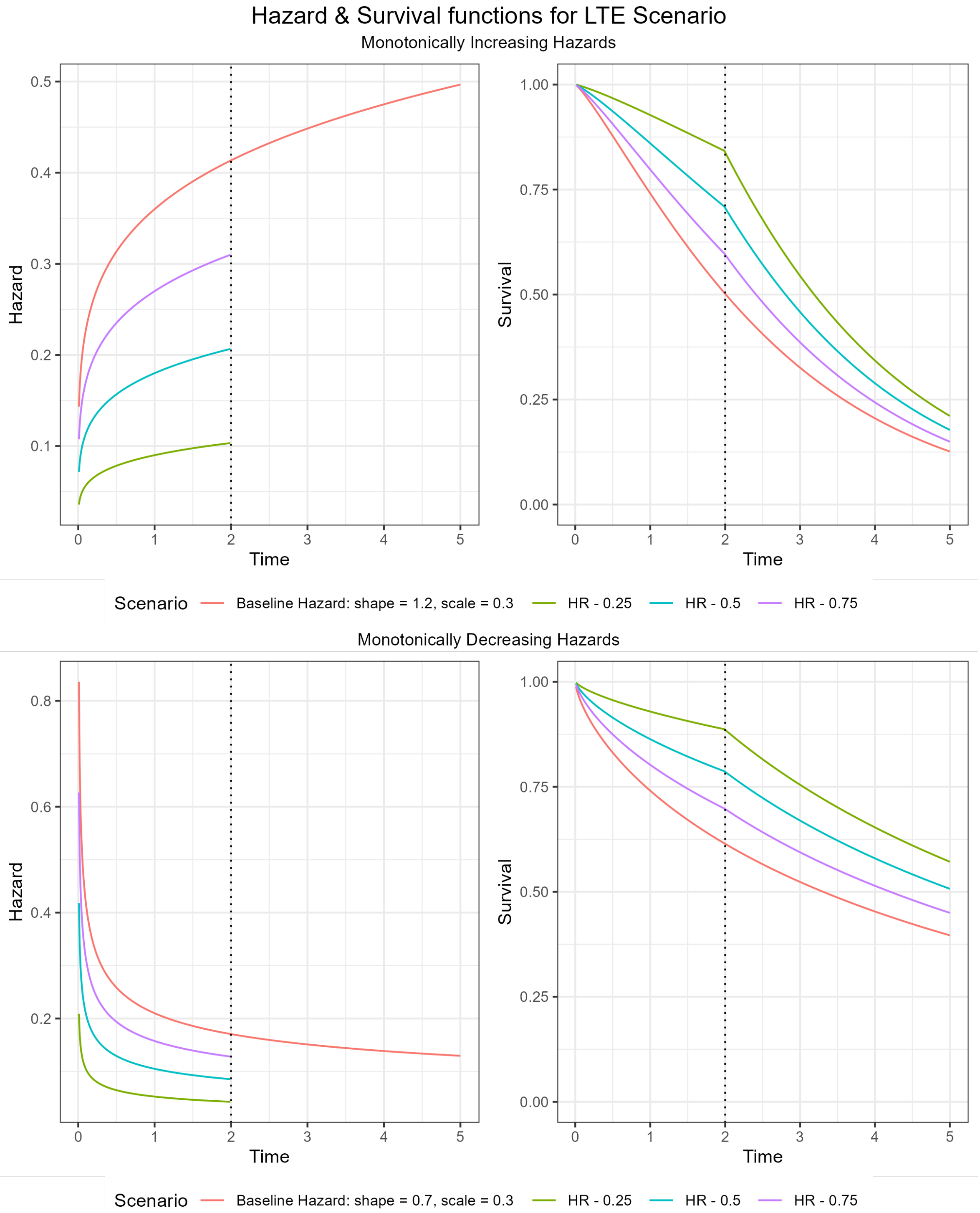


Figure 5.6: Hazard and Survival functions for Loss of Treatment Effect Scenarios

Table 5.2: Simulation Study results - Values of $\text{RMST}_{\text{diff}}$ for Loss of Treatment Effect scenarios

n_{samp}	shape	Initial HR	Change-point	Generalized Gamma	Weibull
200	1.2	0.25	0.08	0.13	0.14
500	1.2	0.25	0.05	0.08	0.14
1000	1.2	0.25	0.04	0.06	0.13
200	0.7	0.25	0.19	0.61	0.58
500	0.7	0.25	0.13	0.57	0.56
1000	0.7	0.25	0.09	0.56	0.58
200	1.2	0.5	0.1	0.16	0.15
500	1.2	0.5	0.06	0.11	0.17
1000	1.2	0.5	0.05	0.08	0.17
200	0.7	0.5	0.31	0.59	0.56
500	0.7	0.5	0.17	0.51	0.53
1000	0.7	0.5	0.14	0.49	0.52
200	1.2	0.75	0.26	0.18	0.18
500	1.2	0.75	0.23	0.14	0.16
1000	1.2	0.75	0.13	0.12	0.17
200	0.7	0.75	0.61	0.5	0.48
500	0.7	0.75	0.44	0.46	0.47
1000	0.7	0.75	0.36	0.46	0.47

observe that the $\text{RMST}_{\text{diff}}$ is lowest for the change-point models with lowest values for scenarios with large sample sizes and smaller values of the HR. A HR of 0.75 produced lower $\text{RMST}_{\text{diff}}$ compared to the next best models for only some of the scenarios, suggesting a relatively large HR is required to adequately estimate the model. For the treatment delay scenarios, the sample size does not impact the $\text{RMST}_{\text{diff}}$ (rounded to two digits). This is possibly because the parameters are more readily identifiable for these scenarios due to the theoretical survival being equal up until the change-point (see Figure 5.5). The theoretical survival function for scenarios assuming equality in hazards after change-points is not equal at any stage and the parameters are possibly more difficult identify.

The results are also sensitive to the baseline hazard. This is because for the scenarios which have a monotonically decreasing hazards result in a higher proportion of the sample being censored at the end of follow up. As a greater proportion of the sample needs to be extrapolated the error in $\text{RMST}_{\text{diff}}$ is also greater. It should be noted that we have not considered the simulation studies for the converging hazards model. This model particularly computationally intensive due to the more complex cumulative function relative to the other change-point models.

For reasons of clarity we did not include all results for the standard parametric models, however, the log-logistic and log-normal models had the worst overall performance across both sets of scenarios. The other parametric models tended to have comparable $\text{RMST}_{\text{diff}}$ and perhaps surprisingly included the simple exponential model. This results from a limitation of evaluating model performance by AUC as an under-estimation of the survival difference before extrapolation could be balanced by an overestimation of survival difference for the extrapolated region. Further analysis could assess $\text{RMST}_{\text{diff}}$ for a restricted set of models which fit reasonably well to the observed data (e.g. including the best four models by certain goodness of fit criteria).

5.5 Examples Datasets

In this section we provide some background on the datasets used and the hypotheses to be tested.

5.5.1 E1690 & E1684 - Multiple Change-point Scenarios

An immunotherapy known as interferon α-2b was evaluated in two observation-controlled Eastern Cooperative Oncology Group (ECOG) phase III clinical trials, E1684 and E1690. The first trial, E1684, was a clinical trial comparing high-dose interferon (IFN) to Observation (OBS). A further confirmatory study, E1690 was initiated in 1991 to attempt to confirm the results of E1684. Various analyses of these trials are presented in Ibrahim et al. (2001).

By combining the E1684 and E1690 (as was also considered in an analysis by Ibrahim et al. (2001)), we obtain a dataset with long-term survival data of a group of patients treated with immunotherapy for multiple myeloma (up to 10 years). This long-term dataset allows us to consider various scenarios, including that the relative treatment effect dissipates, possibly because patients are no-longer receiving treatment. Although not considered in a technology appraisal, this dataset has the previously stated advantage of having a very long-term follow up along with information on potential other covariates of interest which are not available when we digitize published Kaplan-Meier from technology appraisals.

Of interest in a change-point analysis, there is evidence of violation of the proportional hazard assumption for the treatment, but not for other covariates such as age as assessed by Schonefeld residuals. As noted previously, change-point models can investigate a variety of scenarios with respect to the hazard ratio for the treatment effect while also including covariates which do satisfy the proportional hazard assumption.

In the first scenario we consider a Weibull model with a change-point in the hazard ratio for treatment (INF) vs control (OBS). A second scenario considers LTE Weibull change-point model, noting that this differs from the first scenario in that the HR for the second interval is constrained to be equal to 1. We then consider CTE model. The relative goodness of fit of each of these survival models will be presented.

5.5.2 LUME-LUNG 1 - Potential Treatment Delay

In TA347 (2015) the technology of interest was nintedanib + docetaxel vs docetaxel monotherapy in the adenocarcinoma population. Within this population the KM survival functions do not appear diverge until ≈ 5 months. If we fit a standard parametric Weibull (PH) model to the data, the estimated survival assuming proportional hazards may not accurately model the initial section of the data in which the survival functions are very similar. However, a change-point model in which a common Weibull model followed by a Weibull model allowing for a different HR with respect to treatment could allow for a better fit to the data and potentially a more plausible extrapolation.

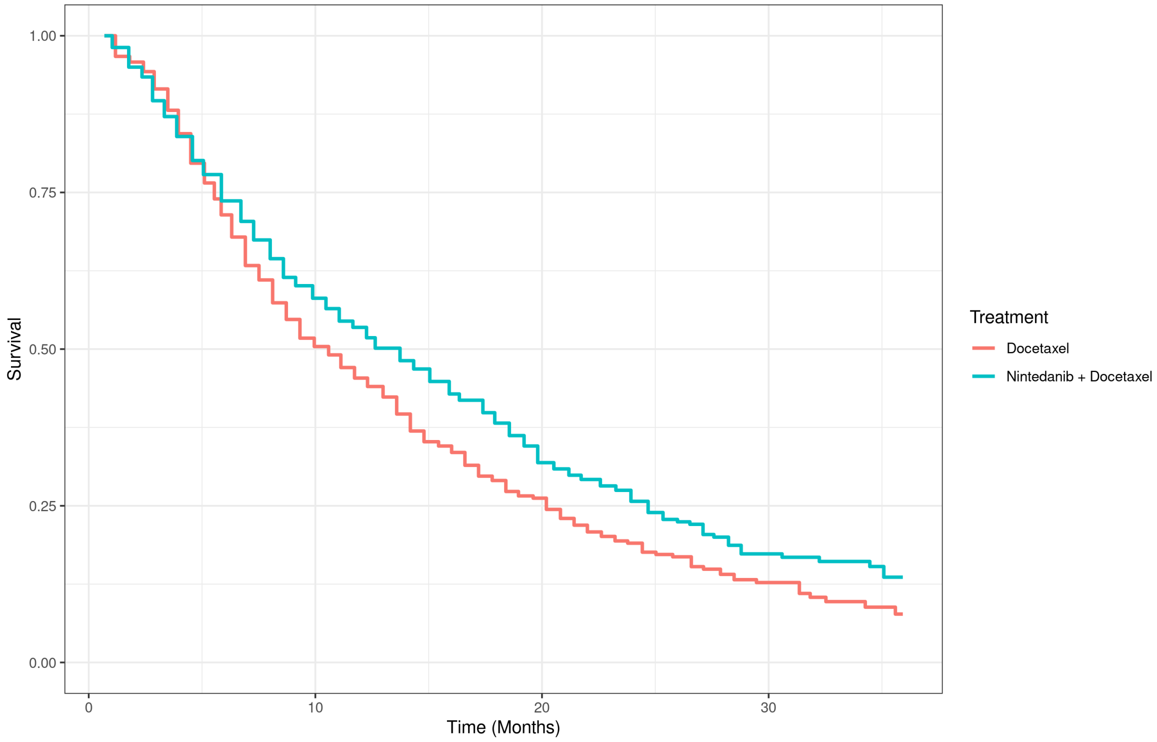


Figure 5.7: LUME Lung 1 trial: Overall survival of adenocarincoma population.

5.5.3 BRIM-3 Study - Loss of Treatment effect

Bagust and Beale (2014) suggest that in the BRIM-3 trial (Chapman et al. (2011)) the effect of vermurafenib is restricted to the first three months of the clinical trial after which a constant common hazard is apparent. They conclude this by inspecting the cumulative hazard plot shown in Figure 5.8 in which they shift the cumulative hazard of the dacarbazine arm by 3 months and note that it approximately lines up with the cumulative hazard of vermurafenib.

hbt!

```
beta_scale
          Interval-1 Interval-2
Intercept       x       x
trt             y       0.0

beta_shape
          Interval-1 Interval-2
Intercept       z       z
trt             w       0.0
```

Listing 5: β covariate matrices for shape and scale parameters in loss of Treatment Effect

It may be of interest to consider whether a constant hazard model or Weibull model fits the data best. We could assess this by fitting two change-point models to the data, one with a constant hazard and another using a Weibull model. In these models, the change-point will apply only to the vemurafenib arm, with the hazard after the change-point estimated from the data beyond the change-point for the vemurafenib arm and the entire data for the dacarbazine arm.

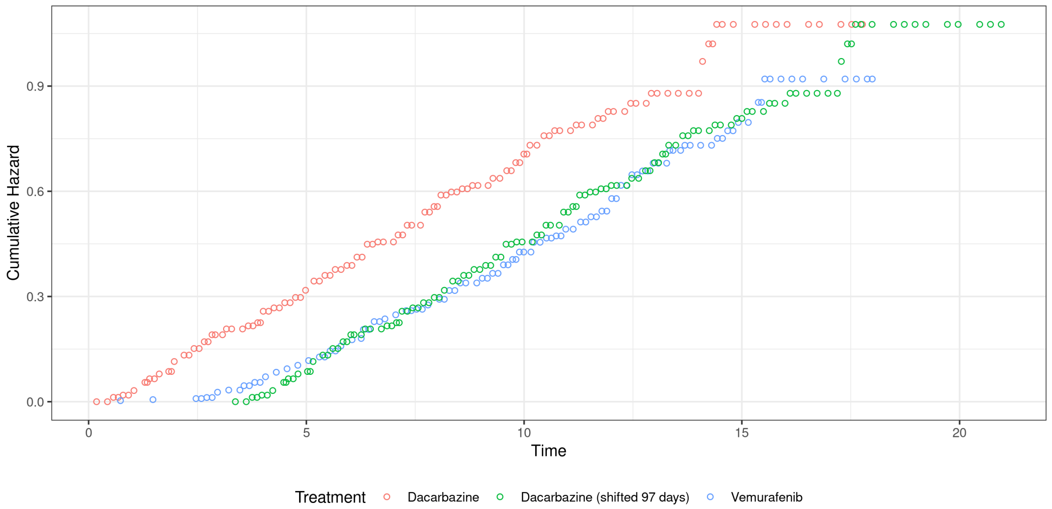


Figure 5.8: BRIM-3 trial: Empirical cumulative hazard plot of overall survival. Reproduced from Bagust and Beale (2014)

.

In terms of the parameters of a change-point model this model can be estimated by allowing a common intercept across intervals for both the shape and scale (See parameters x, z in Code Chunk 5. For interval 1 both the shape and scale are subject to a treatment effect (denoted by parameters y, w), while for interval 2 there is none, denoted by 0. This extends the previous LTE models in that we have a non-proportional hazard model for the first interval then equal hazards after the change-point along with a continuous function for the baseline hazard.

5.6 Model Results

For each of the models estimated below we assume that there is a different baseline hazard function for each interval (i.e. scenario presented in Figure 5.1-B or C) rather than a common baseline hazard function (i.e. scenario presented in Figure 5.1-A). This was because these scenarios provided improved goodness of fit measured by WAIC. In each of the survival functions for the change-point models the posterior density of the change-point was presented (green density with red outline). For each of datasets considered we also fit standard parametric models, calculating the statistical goodness of fit and the $\text{RMST}_{\text{diff}}$ for all models. We also include a Royston-Parmar model which places a treatment effect on the γ_1 parameter allowing for non proportional hazards.

5.6.1 E1690 & E1684 - Various Change-point Hypotheses

For all analysis detailed in this subsection, a covariate for age is included whose HR is fixed with respect to time. In order to plot the survival function stratified by treatment we require the age variable to be set at a particular value (as the survival functions vary with respect to age). In all results presented, hereafter, the survival stratified by treatment is predicted at the mean value of age from the combined trials.

Scenario 1 - Weibull Model step change in HR

In the first scenario we consider a step change in the HR, with the baseline shape and scale parameters also changing before and after the change-point.

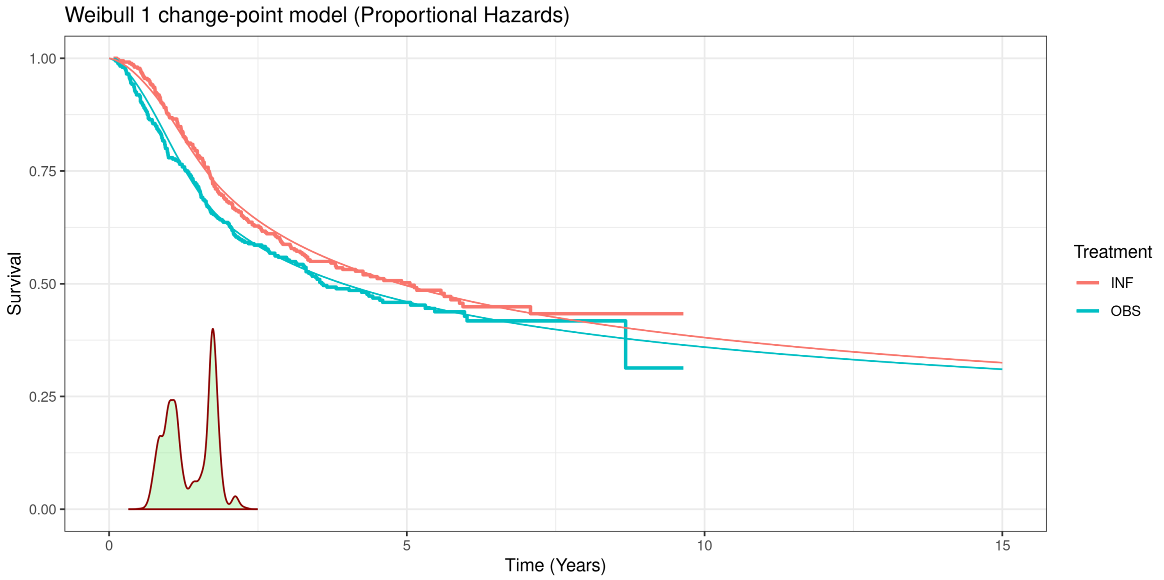


Figure 5.9: E1690 & E1684 trial: Predicted survival function for change-point model for HR with Weibull hazards.

The median change-point is 1.19 years (with a 95% credible interval 0.73-1.90) after which the median HR is 1.07 (with a 95% credible interval 0.8-1.44), suggesting that the treatment effect dissipates (median HR before the change-point was 0.63 with 95% credible interval 0.33-0.93).

Scenario 2 - Common hazards after change-point

In the second scenario we assume that before the change-point there we have a proportional hazards Weibull model and after the change-point the hazards are generated from a common Weibull model with a different baseline shape and scale (Figure 5.10). In this scenario the mean change-point was 1.2 years.

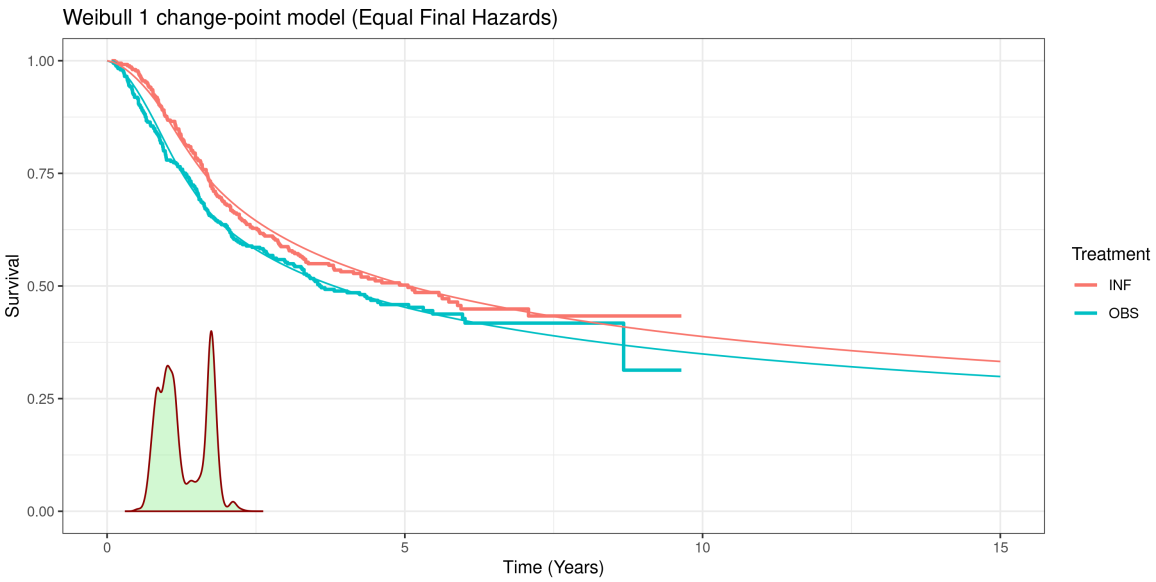


Figure 5.10: E1690 & E1684 trial: Predicted survival function for change-point model with independent then common Weibull hazards.

Scenario 3 - Converging hazards after change-point

A third scenario was the assumption of a converging hazard in which the hazard ratio between the treatment and the intervention converges to a value of 1. The average time of the change-point after which the hazard begins to converge is 1.1 years and the HR before the change-point of had a median value of 0.55 (with 95% credible interval 0.25-0.90).

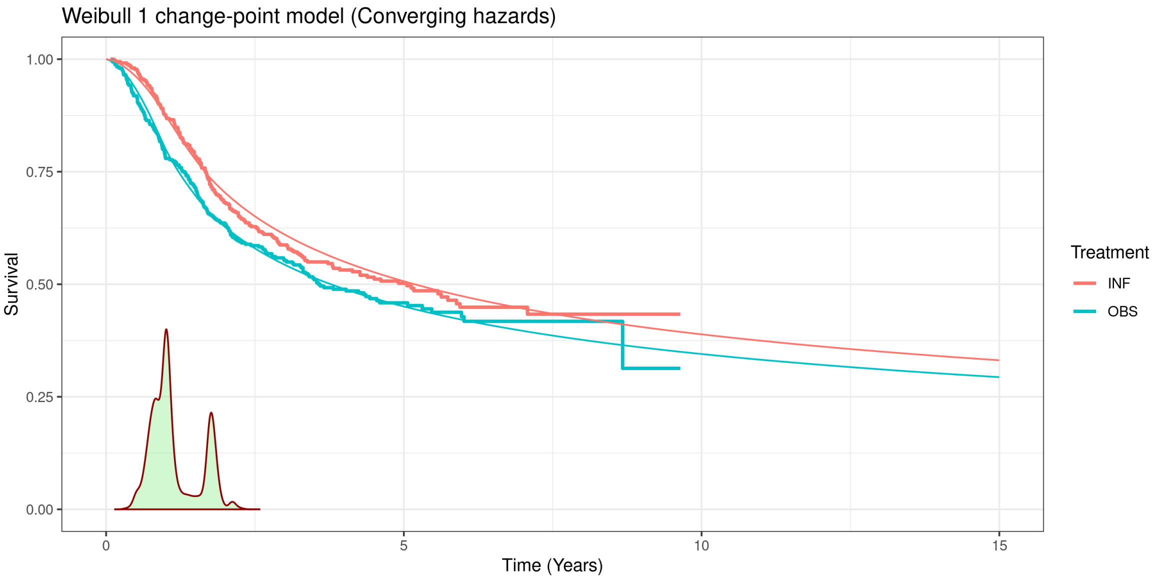


Figure 5.11: E1690 & E1684 trial: Predicted survival function for change-point model with converging hazard ratio over time.

Figure 5.12 shows the posterior distribution of the HR over time and highlights that the HR converges quite rapidly with the median value of the HR converging to 1 before year 2.

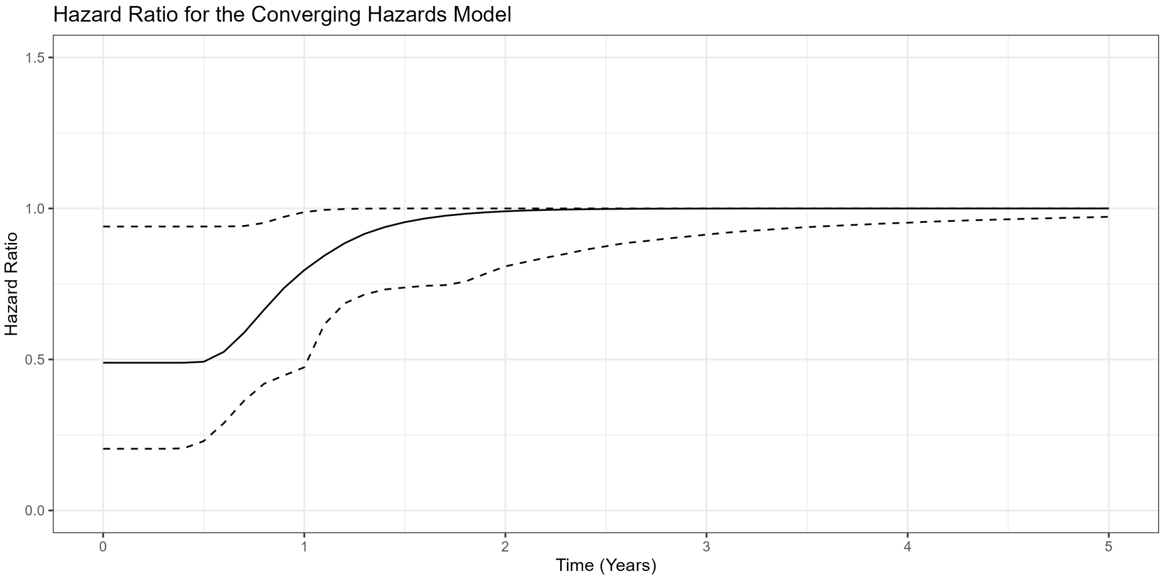


Figure 5.12: E1690 & E1684 trial: Hazard Ratio for change-point model with converging hazards.

Interpretation of results Scenarios 1-3

It is worth noting that each of the methods considered here provide quite similar extrapolated survival estimates, however, the WAIC values are different for the scenario with a change-point for the hazard ratio. The model with the lowest WAIC is the scenario with the converging hazards (Scenario 3 WAIC 2009.18), however, this is effectively equal to the common hazard (Scenario 2; WAIC 2009.35), and then the scenario with a change-point for the hazard ratio (Scenario 1; WAIC 2011.09). The $\text{RMST}_{\text{diff}}$ for Scenarios 2 and 3 are very similar. This is because both models constrain the HR for the extrapolated region to be 1 or very close to 1. It is interesting to note that the model which allowed the HR to be unconstrained for the second interval produced a median posterior HR > 1. This results in the $\text{RMST}_{\text{diff}}$ being lower for this scenario, however, the WAIC indicates that this additional parameter does not produce an improved goodness of fit relative to Scenario 2 (HR constrained to be $= 1$).

It is worth comparing the change-point survival models with flexible models which allow for non-proportional hazards. One such flexible parametric model is the Royston-Parmar cubic spline model which has the option to include "knots" to allow for flexible modelling of the baseline hazard function and which can estimate time-varying hazard ratios by allowing covariates (i.e. treatment status) on the higher-order terms (see `flexsurvspline` from the `flexsurv` package by Jackson (2016) for details). Although the flexible spline model (accommodating non-proportional hazards) visually fits the observed data quite well and has the lowest WAIC at 2003, the survival for the extrapolated region is unlikely to be plausible as there is crossing of the comparator arm (which was observation) with the actively treated arm (at 30 years the survival of the control arm vs the treatment arm is 27% vs 19%). It is worth noting that the standard Royston-Parmar PH model has a WAIC lower than the change-point models and has a similar $\text{RMST}_{\text{diff}}$ (See Table 5.3).

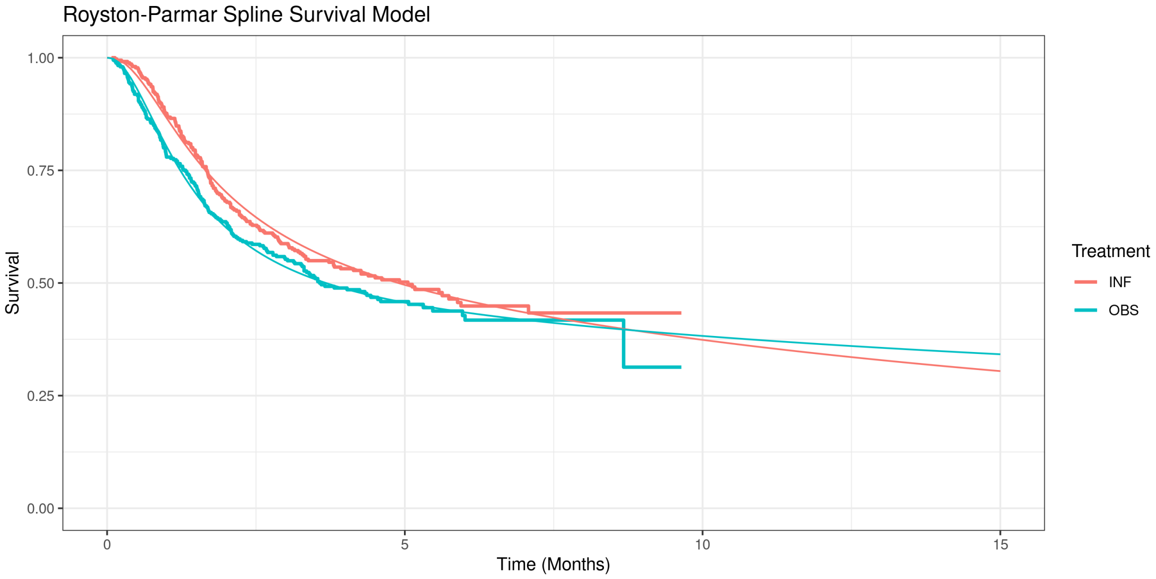


Figure 5.13: E1690 & E1684 trial: Predicted Survival function using Royston-Parmar spline models.

Table 5.3: $\text{RMST}_{\text{diff}}$ for parametric survival models (including all 3 change-point scenarios) and E1690 and E1684 datasets along with WAIC for all parametric models.

Model	$\text{RMST}_{\text{diff}}$	WAIC
Royston Parmar (non-PH)	0.23	2003.03
Royston Parmar (PH)	0.78	2006.70
Change-point: Converging Hazards	0.75	2009.18
Change-point: Equal Final Hazards	0.67	2009.35
Change-point: HR (step)	0.46	2011.09
Generalized-Gamma	1.08	2012.50
Log-Normal	0.93	2036.70
Log-Logistic	0.79	2059.52
Gompertz	0.75	2061.88
Weibull	0.74	2098.73
Exponential	0.74	2100.23
Gamma	0.74	2101.65

5.6.2 LUME-LUNG 1 - Delay of treatment effect

A Weibull model with a change-point assuming a common hazard before the change-point and a separate Weibull model with assuming a proportional hazard model is presented in Figure 5.14.

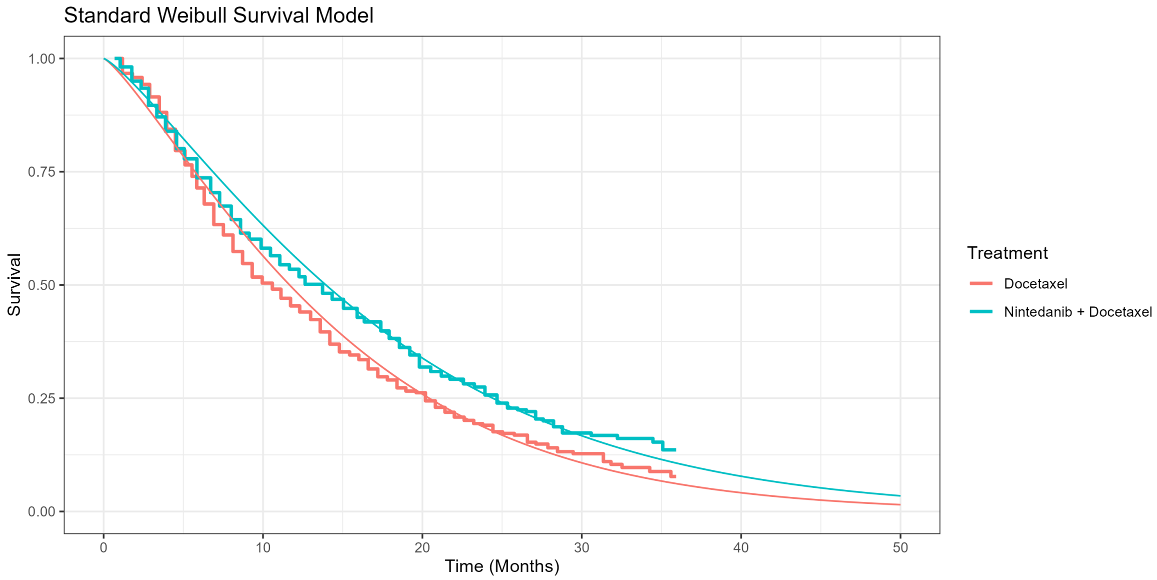


Figure 5.14: LUME Lung 1 trial: Predicted survival from (no change-point) Weibull model

The standard Weibull model appears not to fit the data very well, particularly for the earlier part of the data. Allowing for the a change-point before which the hazards are equal and after which we assume proportional hazards appears to be a better fit to the data (Figure 5.15). In both the standard and change-point model the HR is ≈ 0.8 between the treatment and comparator. The WAIC for the standard parametric model was

3970 while the WAIC for the change-point model was lower at 3933 (Table 5.4).

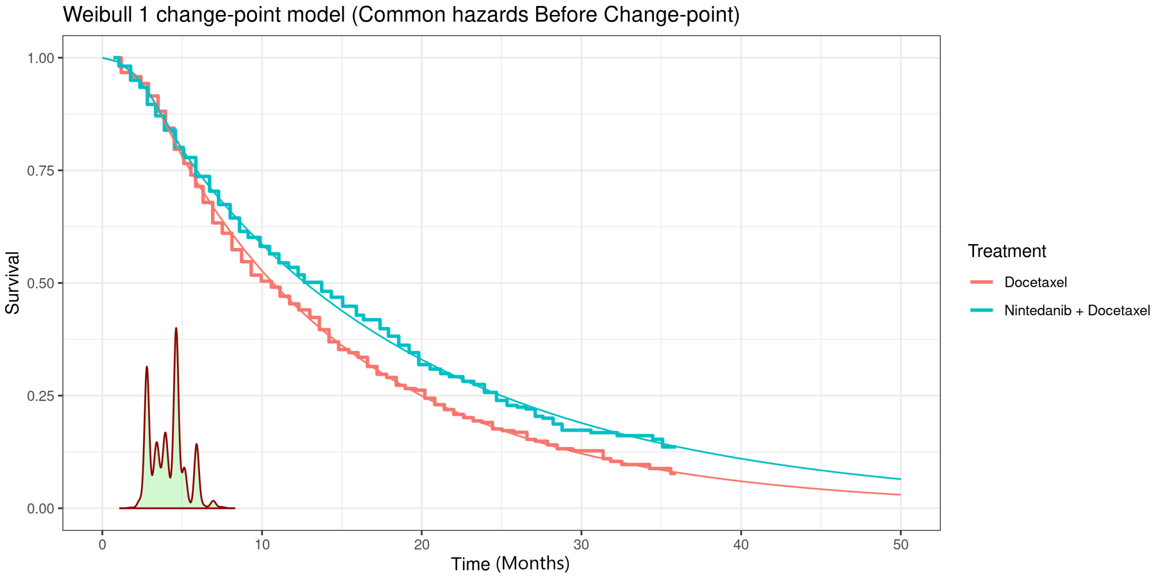


Figure 5.15: LUME Lung 1 trial: Predicted survival from one change-point Weibull model

Table 5.4: $\text{RMST}_{\text{diff}}$ between the parametric survival models and LUME-LUNG-1 along with WAIC for all parametric models - Treatment Delay Scenario.

Model	$\text{RMST}_{\text{diff}}$	WAIC
Change-point	2.72	3933.05
Log-Normal	1.83	3935.17
Generalized-Gamma	1.95	3936.63
Royston Parmar (PH)	2.63	3937.92
Royston Parmar (non-PH)	2.69	3939.61
Log-Logistic	2.09	3945.81
Gamma	2.56	3958.73
Weibull	2.60	3969.97
Gompertz	2.53	3995.38
Exponential	2.51	4001.31

5.6.3 BRIM-3 Study - Loss of Treatment effect

As per Bagust and Beale (2014) we assume a change-point model in which a change-point is considered for the vermurafenib arm and after the change-point the hazard function is equal between vermurafenib and dacarbazine. We consider two change-point models, one assuming constant hazards for each segment compared with a model assuming a Weibull model for each segment.

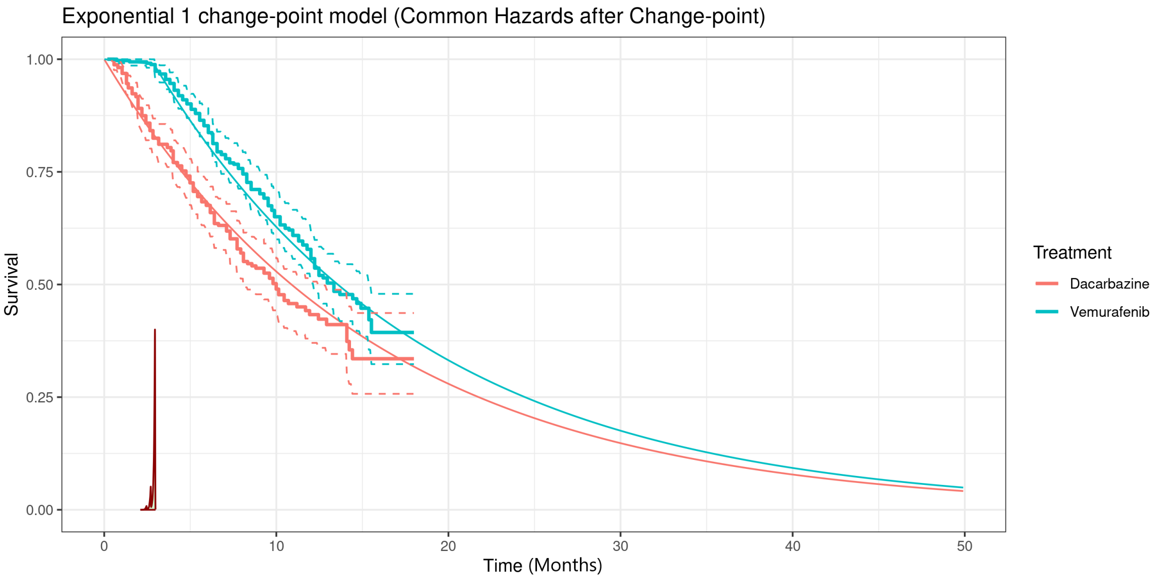


Figure 5.16: BRIM-3 trial: Predicted survival for one change-point model with constant hazards.

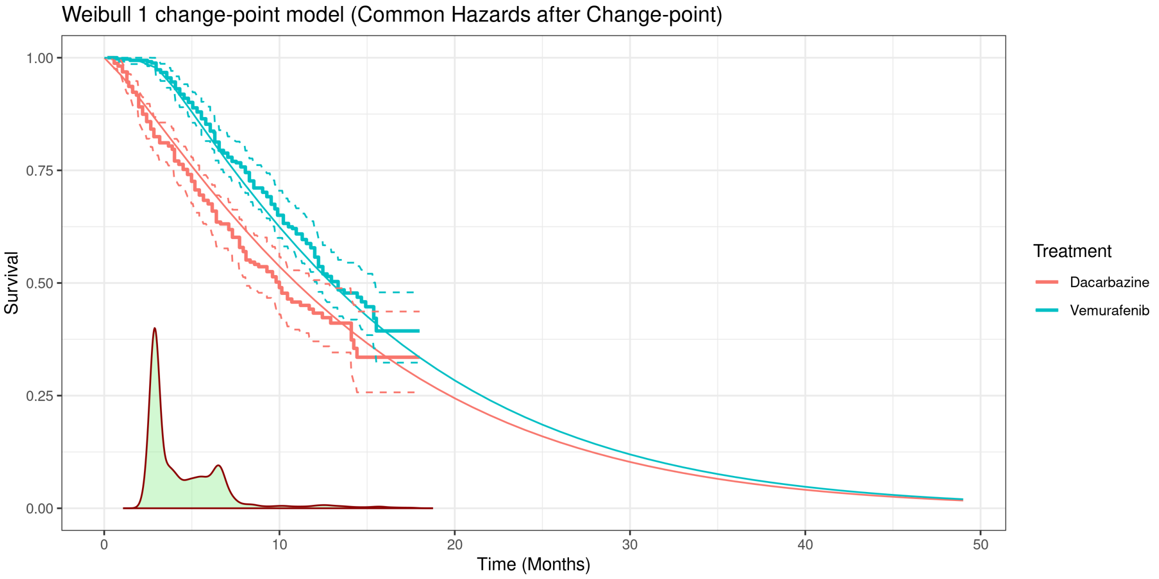


Figure 5.17: BRIM-3 trial: Predicted survival for one change-point model with Weibull hazards.

Comparing this model to the model assuming a hazards generated from a Weibull distribution we see that the survival is very similar between this model and the exponential survival model for the observed portion of the data, however, the extrapolated survival is different (Figures 5.16, 5.17). Because the Weibull model assumes a common monotonically increasing hazard, its survival functions converge more rapidly[1].

[1] In theory the survival functions will only be equal in the limit as $t \to \infty$ so that $S(\infty) = 0$, however, they are practically indistinguishable at 50 months.

Although the models do not exactly match the Kaplan-Meier estimator for the later timepoints it should be noted that the expected survival function remains within the 95% confidence intervals for the Kaplan-Meier trial and have a much lower WAIC than the other parametric models. The WAIC value for both models are very similar and slightly lower for the exponential change-point model, supporting the assertion of constant common hazards as posited by Bagust and Beale (2014). It is worth noting that the posterior distribution for the change-point in the exponential model is very concentrated around 4 months, while for the more flexible Weibull model is much more diffuse. The best fitting model to the data is the Royston Parmar model with non-proportional hazards which fits the observed data well, however, the survival functions quickly cross and the negative $\text{RMST}_{\text{diff}}$ indicates that the over the time horizon (50 months) the expected survival is larger for the dacarbazine arm which is clearly implausible.

Table 5.5: $\text{RMST}_{\text{diff}}$ between the parametric survival models for BRIM-3 along with WAIC for all parametric models - Loss of Treatment effect Scenario.

Model	$\text{RMST}_{\text{diff}}$	WAIC
Royston Parmar (non-PH)	-0.51	2300.18
Change-point Exponential	2.56	2306.92
Change-point Weibull	1.94	2308.60
Log-Normal	6.73	2312.84
Generalized-Gamma	6.84	2314.77
Log-Logistic	5.58	2317.90
Gamma	5.10	2323.74
Royston Parmar (PH)	5.04	2328.00
Weibull	4.54	2330.37
Gompertz	3.38	2351.04
Exponential	5.23	2379.60

5.7 Discussion

In this chapter we have described a general class of survival change-point models and their estimation using modern Bayesian statistical software. Change-point models are particularly useful when modelling data with complex survival functions and when jointly modelling the intervention and comparator in instances when proportional hazards assumption fails. Through simulation studies we have seen that the change-point models produce most accurate extrapolations when the HR between treatment and comparator is substantial along with a large sample size. This is unsurprising as more complex data generating processes require a large number of samples to accurately estimate their underlying parameters. Because change-point models have comparatively more parameters, the associated likelihood surface estimated by these models can be relatively flat, particularly for datasets with high degree of censoring.

Considering real examples, we analysed a large clinical trial dataset with a long follow-up we covariates we considered a variety of scenarios to jointly the model the treatment arms, finding that the relative treatment effect decreased over time. The various approaches to modelling survival produced similar and plausible extrapolated survival, in contrast to the non-proportional hazard flexible spline models which optimized the fit to the observed data but failed to produce sensible extrapolations. We note that the change-point models did not dominate the proportional hazard Royston-Parmar model in terms of goodness of fit, highlighting that the additional complexity in terms of parameters may not provide a true improvement in the fit to the data.

In the LUME-LUNG 1 dataset it appeared from the plot of the empirical survival that the survival did not diverge for a period of time after baseline, however, it is important to assess if the change-point model improved model fit to justify the inclusion of the four additional parameters (baseline shape and scale, change-point and HR for second interval). The final example models the hypothesis that a change-point model is present in one arm followed by common hazard applied to both arms. Previously when Bagust and Beale (2014) suggested a common hazard model, they justified it by visual investigation of the cumulative hazard plots. By considering a change-point model we fully propagate statistical uncertainty while also testing an alternative hypothesis that a Weibull change-point model could have generate the data. We see that assuming Weibull model does have an impact on the extrapolated survival, however, based on goodness of fit it appears the exponential change-point model is most appropriate for the data.

Change-point models developed in this chapter provide a consistent approach to the application of treatment effect waning assumptions which are often a source of disagreement between the company, ERG in technology appraisals (Kamgar et al., 2022). Furthermore uncertainty in the change-point is fully propagated which was a key concern raised by the ERG in TA589 (2019).

The key advantage of change-point models is the flexibility to model a wide variety of scenarios, however, this flexibility does increase the number of potential models . Owing to the presence of potentially many competing scenarios, the modelled hypothesis selected as the basecase should undergo some clinical validation to assess the plausibility of the hypothesis. For example if a common hazard is to be assumed, an expert's opinion may be consulted to assess the timeframe within which this is most probable, and their beliefs formally integrated with the analysis. As mentioned previously change-point models are parameter rich and may be weakly identified from the data. Because of this it is useful to check the posterior distribution of the change-point to see if it has been reasonably informed by the inclusion of data. In contrast a relatively flat posterior distribution for the change-points is suggestive of a weakly informed model.

Rutherford et al. (2020) criticises piecewise models by stating that only the final interval informs the extrapolation. In the case of change-point models while this is also generally true, it does not have to be the case. For example with the converging hazard model the extrapolation is clearly informed by the hazard ratio from the first interval. Additionally we can specify the baseline hazard to be estimated from the entire time interval and only require the hazard ratios for the treatment arm to be interval specific. Even in this situation we could assume a hierarchical model whereby the prior for the hazard ratio for the current interval could be centered on the current value of the hazard ratio for the previous interval (as was considered in a piecewise exponential model with covariates estimated by Ibrahim et al. (2001)).

Additionally Rutherford et al. (2020) consider step changes in the hazard function to be an implausible representation of the disease process. As was demonstrated in this chapter, change-point models need not introduce discontinuities in the hazard function, however, in many situations such models produce a better model fit. As the hazard function is not an observable quantity such as a survival function we don't believe there is a strict requirement for continuity with the function. From our experience with real world survival data, there are many situations in which we observed the empirical survival function dropping precipitously and to attempt to model this with a continuous function may be unrealistic.

5.8 Conclusion

Change-point models could be a useful approach to modelling a variety of hypothesis regarding relative treatment effectiveness. They enable the survival function to be accurately modelled while still allowing enforcing plausible extrapolations for the treatment arms. Although computationally more burdensome than standard parametric models we have shown how these models can be estimated using standard Bayesian software and provide fully worked examples for practitioners to apply to their own datasets. Further research will focus on estimation strategies which improve computational efficiency of these methods and developing a fully functioning R package similar to the `mcp` package by Lindeløv (2020).

As emphasized in this chapter, change-point models allow a high degree of flexibility for modelling both the observed data and the data generating process. As with any parametric model the accuracy of the extrapolation is highly uncertain. If beliefs about the long-term survival of the population are available it is important that these can be integrated into the analysis so that both adequate fit to the observed data and clinically plausible extrapolations are achieved. Integrating these beliefs in a manner which accounts for the uncertainty in the expert opinion relative to information provided by the data is essential. In Part III we develop a methodology to achieve these goals. In Chapter 6 we first consider including expert opinions in general statistical models along with considerations on how to quantify the strength of these opinions. Chapter 7 applies the methodology to parametric survival models including the change-point models described in this chapter.

Part III

Expert Opinion on Observable Outcomes

6 Expert Opinion on Observable Outcomes - A General Framework

6.1 Introduction

In Bayesian analysis there is an explicit allowance for quantitative subjective judgement. However, in a majority of analyses such information is not incorporated (Mikkola et al., 2023). While there are many situations in which expert opinion can be included in a statistical analysis, it is particularly important when data are absent or scarce, and can be used to inform probability distributions or for informing inputs for mechanistic models (Garthwaite et al., 2005). For example, an expert might be asked to specify the expected survival probability of a patient population at a certain time-point, which is unobserved due to censoring, or for their opinion regarding the probability of health care utilization in the future (O'Hagan, 2019). Other applications from fields such as meteorology, agriculture, economics and finance are detailed by O'Hagan et al. (2006).

Given the frequency of situations in which data are unavailable but sensible assumptions are needed, it is reasonable to suppose that expert opinion should be formally included in decision problems utilizing statistical models. However, as noted by Kadane and Wolfson (1998), expertise in a subject-matter is not the same as expertise in statistics and probability, meaning that elicitation of the required inputs often involves training of the experts in statistical concepts and necessitates multiple workshops to gain agreement between experts on a particular input. Consequently, the consensus from the literature is that it is more beneficial to query experts about model observables than model parameters (Kadane and Wolfson, 1998; Garthwaite et al., 2005; Mikkola et al., 2023). In such cases, the underlying elicitation space is the observable space and the form of elicitation can be referred to as "indirect" elicitation. The model observables are variables (e.g. model outcomes) that can be observed and directly measured, in contrast to latent variables (e.g. model parameters) that only exist within the context of the model and are not directly observed (Mikkola et al., 2023).

While considering elicitation on the observable space reduces much of the cognitive burden on the expert, this information must then be encoded on to the parameter space, which is itself a non-trivial task. Among the issues identified by Mikkola et al. (2023) that prevent more widespread incorporation of expert opinion are a) the advancement of a general technical framework applicable to many problems, rather than custom built solutions to specific problems; and b) the lack of good tools for elicitation that integrate seamlessly with existing modelling workflows.

This chapter seeks to address both of these issues within the context of expert opinion on the observable space. We describe a technical framework which can be applied to a broad class of problems. The method can be integrated with commonly used statistical software used for Bayesian analysis or within bespoke code for a particular analysis. We also pay particular attention to quantifying the strength of an expert's opinion, which is an important consideration during elicitation exercises.

The remainder of the chapter is organised as follows. In Section 6.2 we describe existing approaches to incorporating expert opinion on observable quantities. In Section 6.3 we describe the proposed method along with considerations when including expert opinion using the proposed method. Section 6.3.2 highlights approaches to assessing the information contained within expert opinion, measured in terms of an equivalent sample size of data. In Section 6.4 we consider incorporating expert opinion into an exponential model, a relatively straightforward task, which allows us to describe in detail the application of our method and compare it to some of the approaches described in Section 6.2. In Section 6.5 we describe the elicitation of a multivariate normal distribution by using a prior predictive approach in combination with the loss-based framework. The strategy we employ has much fewer and less complex elicitation questions than previous approaches. In Section 6.6 we describe the inclusion of an expert's opinion on mean change from baseline of a treatment in a longitudinal study. To our knowledge inclusion of expert opinion within models analysing repeated measurements has not been considered previously, and we note that our approach is also applicable to generalized linear models.

6.2 Overview of approaches for incorporating Expert Opinion on Observable Quantities

In this section we describe the literature detailing methods for including opinion on observable quantities. This literature falls broadly into four categories, which we briefly describe. While not exhaustive, this categorization covers much of the previous literature.

6.2.1 Prior Predictive Approach

The prior predictive distribution for an unobserved data point x with model parameter θ is defined as

$$\pi(x) = \int f(x|\theta)p(\theta)d\theta,$$

where $p(\theta)$ denotes the prior and $f(\theta|x)$ the sampling distribution for the random variable X. The experts are asked questions about aspects of the data, such as percentiles. Hyperparameters are then estimated so that the prior predictive distribution matches these opinions. The estimation procedure differs between approaches; Percy (2004) finds these parameters through solving systems of non-linear equations, while Wesner and Pomeranz (2021) consider an iterative approach whereby the parameters are manually varied until the prior predictive distribution (evaluated by simulation) appears close to the expert's belief.

Hartmann et al. (2020) ask the expert to provide (prior predictive) probabilities of observables falling in certain regions of the observable space. They then optimize the model hyperparameters, as well as an additional concentration parameter which takes into account that the expert information is itself of a probabilistic nature and hence inherently uncertain.

Prior predictive approaches have been considered when eliciting opinions on observables for linear regression and multivariate normal models (Kadane et al., 1980; Al-Awadhi and Garthwaite, 1998). In these situations some of the hyperparameters were estimated through intermediary calculations by assessing how an expert's belief would change based upon hypothetical data which is different than their current beliefs.

6.2.2 Opinions on value of Response or Percentiles and associated uncertainty

This approach focuses on eliciting opinion on the value of a model response or percentile, and the expert's uncertainty about their estimate, akin to how the uncertainty about a parameter estimate is related through the standard error.

Within this framework, Bedrick et al. (1996) discuss using a class of priors termed data augmentation priors (DAP) to include expert opinion on observable outcomes. DAP have the same form as the likelihood, with the idea that the prior for the parameters is based on "prior observations" that give rise to a likelihood that has the same form as the likelihood for the data. This results in a posterior that has the same form as the likelihood. Bedrick et al. (1996) and Johnson (1996) detail how to define such priors for generalized linear models (GLMs) and survival data modelled by a log-normal distribution respectively.

Hosack et al. (2017) provide a method for including expert opinion on the expected response of a GLM by assuming that the $\boldsymbol{\beta}$ regression coefficients from the linear component are multivariate normal ($\mathcal{MVN}$). They then minimize the Kullback-Leibler divergence between the percentiles of the expected response from the $\mathcal{MVN}(\beta)$, transformed using the appropriate link function, and the expert's percentiles.

6.2.3 Parameter Reparameterization

If a model can be reparameterized so that the observable quantity in question becomes a parameter in the model itself, then the expert's opinion can be included directly on this parameter. Singpurewalla and Song (1988) consider expert opinion about median survival modelled by a Weibull distribution, which was reparameterized to include a parameter for median survival and (the standard) shape parameter. The expert is asked to provide a mean and variance for median survival and then an informative prior for the shape parameter. Assuming independence of these parameters, a joint prior is defined which can then be updated with data in the standard fashion. Although allowing the expert to define the first and second moments of the median survival, the distribution of median survival is constrained to be a function of a chi-squared distribution and the specification of a prior for the shape requires elicitation on the parameter space.

Wongnak et al. (2022) consider a Weibull distribution with expert opinion on mean survival. In this hierarchical approach, mean survival was treated as a model parameter, with the scale parameter a deterministic function of mean survival and the shape parameter. The mean survival is assumed to have a log-normal distribution whose parameters were elicited from the experts, while the shape parameter is assigned a vague uniform prior.

6.2.4 Eliciting Opinion based on imagined data observations

Another methodology which has received consideration from both Bayesian and frequentist perspectives is to augment knowledge about the process under study by the use of data elicited from experts. For statistical models with fixed dimension sufficient statistics, it can be possible to elicit opinions in terms of data. For example, in the case of a binomial likelihood, a Bayesian who asserts that their opinion is equivalent to observing seven out of ten successes has provided a $\mathcal{B}(7,3)$ beta distribution, while in the frequentist analysis the "data" enters the likelihood as if they were real observations.

See Coolen (1996) for some examples on setting hyperparameters based on imagined data in survival models. Low Choy et al. (2013) discuss incorporating expert opinion on observables for a logistic regression model using a DAP but also note that the opinions could be included through the addition of pseudo-observations.

Lele and Das (2000) use hierarchical models in a frequentist framework to formulate the problem of combining observed data and guess values obtained from experts using a credibility parameter (a form of calibration), which assesses the correlation between the expert's guesses and known values.

6.2.5 Summary

While there are a number of strategies to including expert opinion on observables, it is clear that none of the approaches are truly model agnostic. That is, they do not have the ability to incorporate any and all types of expert opinion on observable outcomes with all possible forms of likelihood. While the reality is that it is unlikely that such an approach exists, it is clear that some methods are easier to generalize than others.

Prior predictive approaches are specific to the prior and likelihood combination, and for likelihoods with multiple parameters, numerical techniques may be required to obtain hyperparameters (Percy, 2004; Hartmann et al., 2020). Although prior predicitive approaches for obtaining expert opinion on multivariate normal models and linear regressions are mathematically elegant, they are quite cognitively burdensome on the expert. This is discussed further in Section 6.5.

The approaches of Bedrick et al. (1996) and Hosack et al. (2017) offer solutions for an important class of statistical models, GLMs. An advantage of the approach of Hosack et al. (2017) is that the computational approach used to induce priors for different types of GLM is the same, and is implemented as an R package. This is in contrast to Bedrick et al. (1996), who require model specific calculations to derive the induced priors.

The parameter reparameterization used by Wongnak et al. (2022) is generalizable to other statistical models. However, there are many examples where an observable outcome cannot be expressed analytically in terms of the other parameters. For example, the mean survival of a Gompertz distribution involves an intractable integral.

Including expert opinion by imagining data allows for the straightforward incorporation into the statistical analysis. However, it can also require restrictive assumptions. For example, in Coolen (1996), all of the parametric models with more than one parameter were assumed to have the second parameter fixed.

6.3 Loss based approaches to incorporating expert opinion

An alternative approach to those described in Section 6.2 is to adapt the framework of Bissiri et al. (2016), in which expert belief can be incorporated through the use of a loss function. In this section we introduce the required notation, highlight the types of expert opinion which can be included using this framework and discuss practical considerations when eliciting opinions and implementing an analysis.

6.3.1 Loss Adjusted Posterior approach

A valid and coherent update of a prior $p(\theta)$ to a posterior $\pi(\theta|x)$ through a (negative) exponentiated loss function (Bissiri et al., 2016) is

$$\pi(\theta|x) \propto \exp\{-l(\theta, x)\}p(\theta). \tag{6.1}$$

Importantly this is the standard Bayesian update if the loss function is the negative log-likelihood $l(\theta, x) = -\log\{L(x|\theta)\}$. We propose to incorporate expert opinion into a model by specifying a loss function that includes the parameters of a probability distribution ϕ that describe the expert's opinion about the observable quantity. This distribution is itself a function of the model parameters, $g(\theta)$. This idea extends the framework established by Bissiri et al. (2016), who considered opinion on the parameter space rather than the observable space. Within this proposed framework, we replace x with ϕ and θ with $g(\theta)$ so that the loss function becomes

$$\pi(\theta|\phi) \propto \exp\{-l(g(\theta), \phi)\}p(\theta). \tag{6.2}$$

A key point to highlight is that the focus of this approach is to construct a posterior distribution which directly encodes the expert's beliefs about the observable space. This is distinct to the methods described in Section 6.2, where the emphasis is on the model prior and the identification of suitable hyperparameters that indirectly describe the expert's opinion on the observable space. This loss-adjusted posterior can be naturally updated when data are available using the model likelihood, resulting in a posterior including both expert opinion and data:

$$\pi(\theta|x, \phi) \propto \exp\{-(l(\theta, x) + l(g(\theta), \phi))\}p(\theta). \tag{6.3}$$

The posterior distribution for the model parameters is hence a function of both the fit to the observed data and to the observable quantity defined by the expert. We abbreviate the approach described above as LAP (Loss Adjusted Posteriors) which highlights that a particular prior is updated with a loss function to give a posterior which includes the expert opinion.

As we will show through examples in the subsequent sections, this approach allows us to be very flexible with respect to the types of information we wish to include in the statistical analysis. Importantly, it is straightforward to include the loss function within existing Bayesian software such as Stan (Stan Development Team, 2020) or JAGS (Plummer, 2003). In Stan, it is possible to increase the log density of the posterior by employing a specific statement in the Stan syntax denoted as `target +=`. This statement allows for the target value to be incremented by a specified amount. For instance, in Code Chunk 6 of Section 6.4.3, we utilize the `target +=` statement, where the right-hand side of the operation corresponds to the negative of the loss function $-l(g(\theta), \phi)$.

The outlined LAP approach can only incorporate opinions about observable quantities which can be expressed as deterministic functions of the model parameters, such as the mean, variance or percentiles, and so on. We cannot specify a loss function which encodes a belief on the percentiles of a prior predictive density, which integrates out the parameter(s) of interest.

Two important points are worth emphasising regarding this approach. Firstly, the choice of prior $p(\theta)$ will impact the posterior $\pi(\theta|\phi)$, which potentially could result in the posterior density for the observable quantity being different to that imagined by the expert. Typically a vague prior will have a minimal impact on the resulting LAP, however, as described in Section 6.4 it is possible to remove this impact entirely by defining the observable quantity as a parameter (typically with a uniform prior) so that the LAP produces a density of the model observable which exactly matches the expert's opinion. Although this is operationally similar to the hierarchical approach described in Section 6.2.3, the LAP approach is more general and can include expert opinion when parameter reparameterizations are not feasible.

Secondly, we must ensure that the resulting LAP is a proper distribution, meaning that the integral over the parameter space is finite. Fitzpatrick (2009, Chapter 6) demonstrates that a continuous function over a finite interval $[a, b]$ is integrable, meaning that it evaluates to a finite number. Since we require a finite interval, we specify proper priors in our applied examples.

For the exponential likelihood in Section 6.4, we establish the propriety of the posterior distribution when incorporating a loss function that models an expert's belief on median survival as a log-normal distribution. Because the exponential model is univariate, it enables us to explore different combinations of loss functions and priors, determining whether they yield proper or improper posteriors. These illustrative examples offer valuable insights into the conditions necessary for a proper posterior, emphasizing the importance of having a proper prior. For examples in Sections 6.5 and 6.6 we define proper priors and assume that inclusion of the loss functions do not introduce discontinuities in the posterior distributions of the parameters, therefore, that the posterior distributions are proper.

As discussed by Bissiri et al. (2016), when using a loss based framework we are free to multiply either loss by an arbitrary factor which we denote by weights w_1, w_2 for the losses for the data and expert opinion respectively (Equation 6.4).

$$\pi(\theta|x, \phi) \propto \exp\{-(w_1 l(\theta, x) + w_2 l(g(\theta), \phi))\} p(\theta). \tag{6.4}$$

For the examples considered in this chapter, the loss with respect to the data is the special case of the negative log-likelihood function so it is reasonable to set $w_1 = 1$ (as in the standard Bayesian update). Our opinion is that the weight on the loss for the expert opinion should typically be set to 1. This is because the strength of the expert's belief is already governed by the parameters for the distribution representing their belief. If the expert has a stronger belief they (or the elicitation process) will naturally calibrate the parameters of the distribution representing their belief to have a higher precision, which can be quantified by the Effective Sample Size (ESS) in Section 6.3.2 rather than adjusting w_2. If however, the analyst believes that the expert's opinion is still overconfident it may be useful to calibrate $w_2 < 1$ which essentially increases the variance of the distribution associated with the belief. In all subsequent examples in the chapter we assume $w_1 = w_2 = 1$. The consideration of weights w_1, w_2 has a relation to power priors/posteriors discussed by Ibrahim et al. (2015), where the likelihood is raised the the power of a value a_0 which can be between $[0, 1]$. They state that a power prior will be proper if the initial prior is proper. We make the assumption that if the posterior distribution associated with the loss function $l(g(\theta), \phi)$ is continuous, then the posterior associated with the power of the same loss function is also continuous, and consequently, integrable.

One other consideration relates to conjugate posteriors. Considering the update implied by Equation 6.1, including expert opinion implies that except in very special circumstances[1] the model will not be conjugate. This is more of a historical concern as software such as Stan does not rely on conjugacy and while JAGS can use conjugacy for computational efficiency it can alternatively use slice sampling when this option is not available (Bølstad, 2019).

[1] One trivial example would be to include expert opinion on the expected probability of success as a Beta distribution with a Beta prior and a binomial likelihood.

6.3.2 Quantifying the strength of expert opinion

In any analysis which includes subjective beliefs or external information, it is important to quantify the strength of the supplied evidence. This is particularly true when eliciting expert opinion on quantities in either the parameter or observable space, as apparently modest changes in opinion can be associated with a large change in the estimated precision of the model parameters.

One active area of research for quantifying the strength of an opinion is the definition of the effective sample size (ESS), denoted as n_e. Conventionally, this quantity defines the amount of information (in terms of n observations) contained in the prior (Morita et al., 2008). Alternatively, the term can define the difference between a posterior incorporating an informative prior and a posterior using a reference/vague prior (Reimherr et al., 2021). One important distinction between these methods is that posterior-based approaches can potentially account for prior-data conflict, which can in theory produce negative ESS values.

The ESS of $\pi(\theta|\phi)$, the expert-informed loss function component of the LAP, can be quantified in the same manner as the prior $\pi(\theta)$ in the framework of Morita et al. (2008), if we can similarly assume that this density is approximately normal. Alternatively, the LAP component containing expert opinion and data, $\pi(\theta|x, \phi)$, can be compared to a standard posterior with a vague prior $\pi(\theta|x)$, similar to the method of Reimherr et al. (2021). In the examples that follow we consider different approaches to approximate the ESS, with the pragmatic aim of gaining a reasonably quantified feeling of the impact of the external information, e.g. whether it is 5% or over 30% of the data we have collected, and not attempting to pinpoint the exact contribution (Reimherr et al., 2021).

6.4 Exponential Likelihood

In the following sections we incorporate hypothetical expert opinion on percentiles using the prior predictive approach, and also by eliciting an expert's belief about a percentile and their associated uncertainty using DAP and LAP approaches. This section highlights the different information that is required for these approaches using an exponential model for which inference is relatively simple.

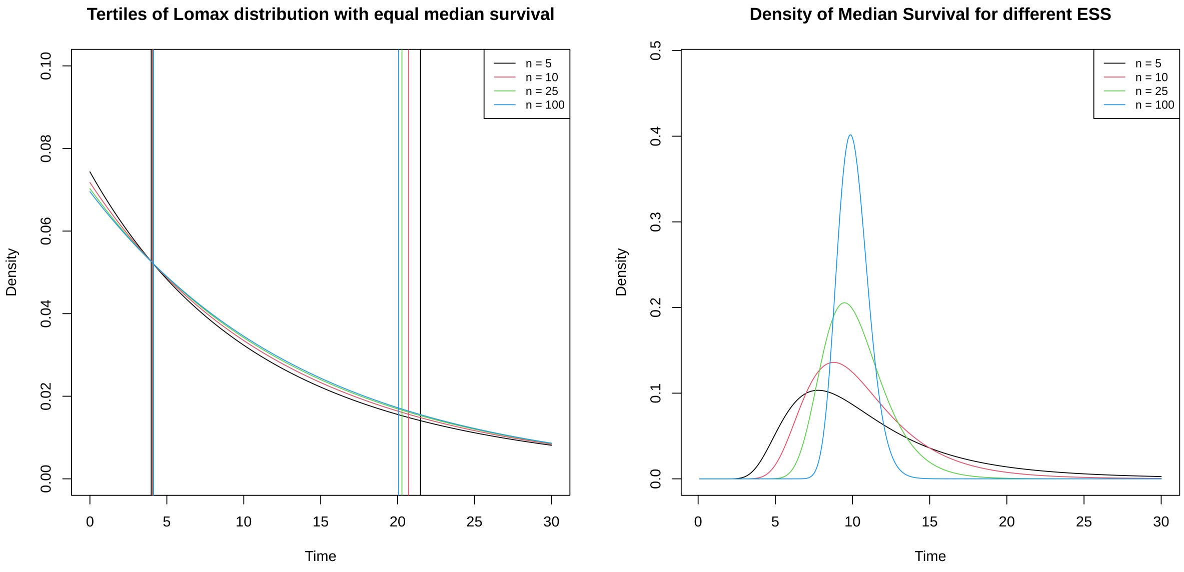
The exponential model is presented in two separate parameterizations; $y \sim \mathcal{E}xp(\lambda)$ or $y \sim \mathcal{E}xp(1/\psi)$ with y the survival time of an observation. The hazard is denoted by λ and $\psi = 1/\lambda$, with ψ also the mean survival time of the observations. For the prior predictive approach we have a gamma distribution as the prior for $\lambda \sim \mathcal{G}(\alpha, \beta)$, while we have an inverse-gamma distribution for $\psi \sim \mathcal{IG}(\alpha, \alpha\tilde{y})$ or alternatively $\beta = \alpha\tilde{y}$. For the LAP method we consider both parameterizations, with their respective priors and loss functions discussed in detail.

6.4.1 Prior Predictive Approach

Percy (2004) considers asking experts to specify quantiles by stating two times, $Q_{1/3}$ and $Q_{2/3}$, such that the lifetime of an object was equally likely to be in each of these three intervals: $[0, Q_{1/3})$; $[Q_{1/3}, Q_{2/3})$;$[Q_{2/3}, \infty)$. Of course we are not restricted to specifying quantiles and could elicit any two percentiles, e.g. $Q_{1/2}$ and $Q_{3/4}$ representing the 50th and 75th percentiles. Assuming a gamma prior and an exponential sampling distribution, the prior predictive distribution is know as the Lomax distribution, with cumulative distribution function $F(x) = 1 - (\frac{\beta}{x+\beta})^{\alpha}$. It is then straightforward to find the parameters which satisfy the values of $Q_{1/3}$ and $Q_{2/3}$.

While analytically tractable, in practice it can be challenging to encode an expert's uncertainty into the analysis using this approach. Figure 6.1 shows the $Q_{1/3}$ and $Q_{2/3}$ tertiles of four different gamma distributions which have the same median survival but different levels of ESS, $n_e = 1, 10, 25$ and 100. Also shown is the prior density for the median for each of these samples sizes. In this case ESS is simply $n_e = \alpha$, owing to the properties of conjugacy. It may not be clear to the expert why their opinion becomes more or less informative by changing the values of $Q_{1/3}$ and $Q_{2/3}$, especially as small changes in the percentiles can result in large changes in the informativeness of the prior. One potential solution could be to elicit only one percentile (e.g. the median) and also elicit an effective sample size.

It is also challenging to extend this approach to alternative sampling distributions other than the exponential. Percy (2004) discusses but does not implement an approach for the Weibull distribution, for which there is no available analytic expression for the prior predictive distribution. This approach then requires numerical methods for both the evaluation of the prior predictive distribution and identification of the hyperparameters, of which four are required.



(a) Quantiles of Lomax distributions representing different levels of ESS.

(b) Median Survival representing different levels of ESS.

Figure 6.1: Expert Opinion for exponential likelihood using Prior Predictive Approach

6.4.2 Data Augmentation Prior Approach

Bedrick et al. (1996) discuss DAP for an exponential likelihood including expert beliefs at different values of a covariate, however, for the purpose of comparison we will specify data augmentation priors for the intercept only model. For this model the DAP is constructed as $\psi \sim \mathcal{IG}(\alpha, \alpha\tilde{y})$, where α is the sample size and $\tilde{y}$ is a parameter representing the mean survival time and ψ represents a random draw of mean survival.

We can specify percentile(s) for the median survival $t_{\text{med}} = \psi \log(2)$ and, after the appropriate calculations for the change of variables, relate opinion about median survival to the inverse gamma distribution. Equation 6.5 provides the density on the median survival, where for clarity $\beta = \alpha\tilde{y}$:

$$f(t_{\text{med}}|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \left(\frac{t_{\text{med}}}{\log(2)}\right)^{-\alpha-1} \exp\left(\frac{-\beta \log(2)}{t_{\text{med}}}\right) \frac{1}{\log(2)}. \tag{6.5}$$

Similar to the example in Section 6.4.1 we could ask experts to specify quantiles $Q_{1/3}$ and $Q_{2/3}$ and find parameters α, β through optimization. Importantly, these quantiles represent the quantiles of median survival and not the data distribution. In this case, the quantiles represent the expert's uncertainty around the value that median survival takes and narrower intervals naturally represent more informative priors.

In Figure 6.2 we see the resulting expert belief on median survival after they have specified the 50th percentile for the median as 10 and 75th percentile as 13.85. This produces the same parameters for α and β as assuming lower and upper quantiles of 3.98 and 21.49 for the Lomax distribution (i.e. n = 5 in Figure 6.1 (a)).

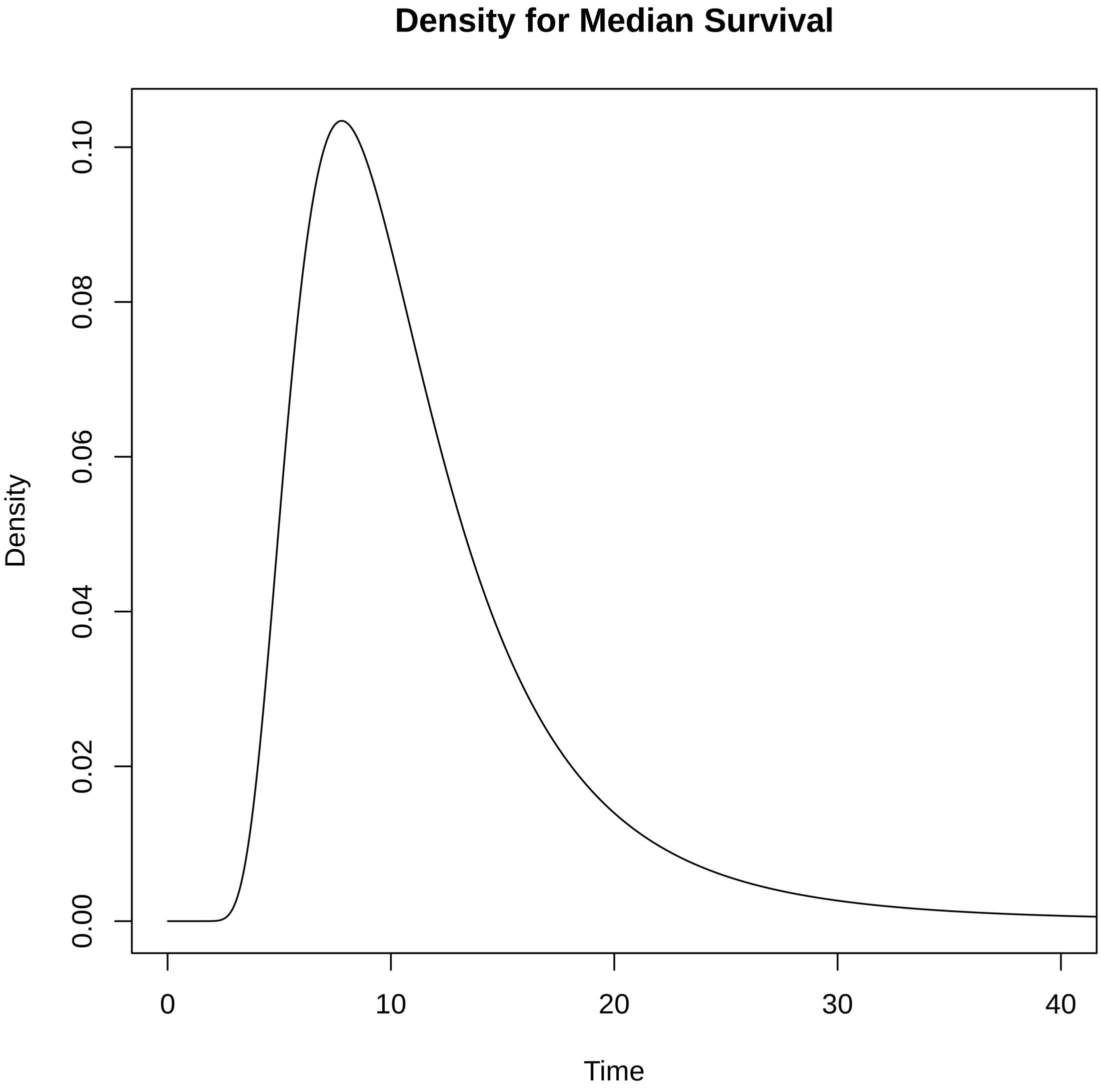


Figure 6.2: Prior on Median Survival based on $\mathcal{IG}(5, 67.25)$ prior

6.4.3 Loss Adjusted Posteriors

In Section 6.4.2 we have applied the DAP approach to incorporate expert opinion on median survival. However, we note two primary limitations. Firstly, these priors are limited to a number of special cases, as described by Bedrick et al. (1996) and Johnson (1996), however, even in these cases deriving the model hyperparameters typically involves non-trivial calculations. Secondly, in some situations, the expert may wish to not only provide the percentiles but may also wish to specify a lepokurtic distribution to provide a degree of robustness to their opinion or possibly specify a non-parametric histogram prior (O'Hagan et al., 2006).

One approach which avoids these disadvantages is to incorporate information on the model observables through a loss function. One clear advantage is that the expert is not restricted in the distributions that they can use to describe their belief about the observable quantities. For illustration we suppose that the expert would like to represent their belief about median survival using a log-normal distribution. This is straightforward to encode when using a loss function. For the purpose of illustration we consider a log-normal distribution with the same mean and variance as that presented in Figure 6.2, which through method of moments has the parameters $\mu_{\text{expert}} = 2.31$ and $\sigma_{\text{expert}} = 0.53$. We will incorporate this opinion under both parameterizations of the exponential model i.e. ψ and λ.

Expert Opinion under ψ parameterization

To compute the loss function we express the model parameter ψ in terms of the observable quantity which for the median survival and exponential likelihood is $g(\psi) = t_{\text{med}} = \psi \log(2)$. According to the expert's belief this opinion is log-normally distributed $\mathcal{LN}(\mu_{\text{expert}}, \sigma_{\text{expert}})$ and the loss function comprises of the deviation of t_{med} generated by ψ to the expert's belief. Therefore the loss function $l(g(\psi)|\mu_{\text{expert}}, \sigma_{\text{expert}})$ encodes this contribution as $-\log \mathcal{LN}(\psi \log(2)|\mu_{\text{expert}}, \sigma_{\text{expert}})$, amounting to a loss function

$$l(g(\psi)|\mu_{\text{expert}}, \sigma_{\text{expert}}) = \log(\psi) + \frac{1}{2}\left(\frac{\log \psi - \mu_{\text{expert}}}{\sigma_{\text{expert}}}\right)^2.$$

If we define the loss contribution as above and specify a uniform prior for ψ, then we can generate samples from $\pi(\psi|\phi)$ using e.g., Markov chain Monte Carlo (MCMC), then in the limit the posterior distribution for ψ produces a median survival which will exactly match the expert's opinion encoded as an $\mathcal{LN}$ distribution for t_{med}.

Expert Opinion under λ parameterization

Similar to the previous section we express λ in terms of the median survival for the exponential model which is $g(\lambda) = t_{\text{med}} = \log(2)/\lambda$. The expert's contribution to the loss function is $-\log \mathcal{LN}(\log(2)/\lambda|\mu_{\text{expert}}, \sigma_{\text{expert}})$. However, if we place a uniform prior on λ, the posterior distribution for t_{med} will not be exactly $\mathcal{LN}$. This is because setting a uniform $\mathcal{U}$(a, b) prior on λ results in a non-uniform prior on the median survival, $p(t_{\text{med}}) = \frac{1}{b-a}\frac{\log(2)}{t_{\text{med}}^2} = \frac{1}{b-a}\frac{\lambda^2}{\log(2)}$, and therefore contributes information to t_{med} in addition to the expert's opinion.

Because we have a closed form expression for the density of median survival implied by a uniform prior on λ we can include (a function of) this density explicitly in the loss function. This essentially cancels out the prior contribution of the uniform prior for λ on the median survival. The final expression for the loss function in terms of λ is then:

$$l(g(\lambda), \phi) = -\log(\mathcal{LN}(\log(2)/\lambda|\mu_{\text{expert}}, \sigma_{\text{expert}})) + \log\left(\frac{1}{b-a}\frac{\lambda^2}{\log(2)}\right). \tag{6.6}$$

Using this loss function to update a uniform prior on λ will give a posterior distribution for λ which which has a log-normal median survival, i.e., the LAP. In most multivariate examples it will not be possible to derive a closed form expression of the density of the observable outcome implied by the prior for the parameters. In order to deal with such situations we can typically reparameterize the model so that the observable quantity is a parameter with a prior distribution (typically uniform) and one of the model parameters is a function of the observable quantity (and the other model parameters). In this hierarchical model specification, the model parameter is a logical or deterministic function observable quantity (and the other model parameters).

Once the technical conditions ensuring the propriety of the update are met, a key strength of this approach lies in its straightforward implementation. This simplicity is exemplified by the pseudocode for the Stan model, as presented in Code Chunk 6, which incorporates expert opinion on median survival (Figure 6.3). In this example the upper and lower bound of the uniform prior for λ (i.e. a, b is 0.001 and 10 respectively).

```
transformed parameters{
//lambda is a deterministic function of median St
  median_St = log(2)/lambda;
}
model{
 lambda ~ uniform(0.001, 10);
 target += lognormal_lpdf(median_St|mean_expert, sd_expert) - log(lambda^2)
}
```

Listing 6: Pseudo Stan code implementing expert opinion. The final line in this example uses the `target +=` statement to increment the typical model posterior to include the additional loss function,omitting constants which do not involve model parameters.

We could even implement the expert's belief non-parametrically using the histogram method or parameterize the expert opinion as a (truncated) Student's t distribution with a low number of degrees of freedom.

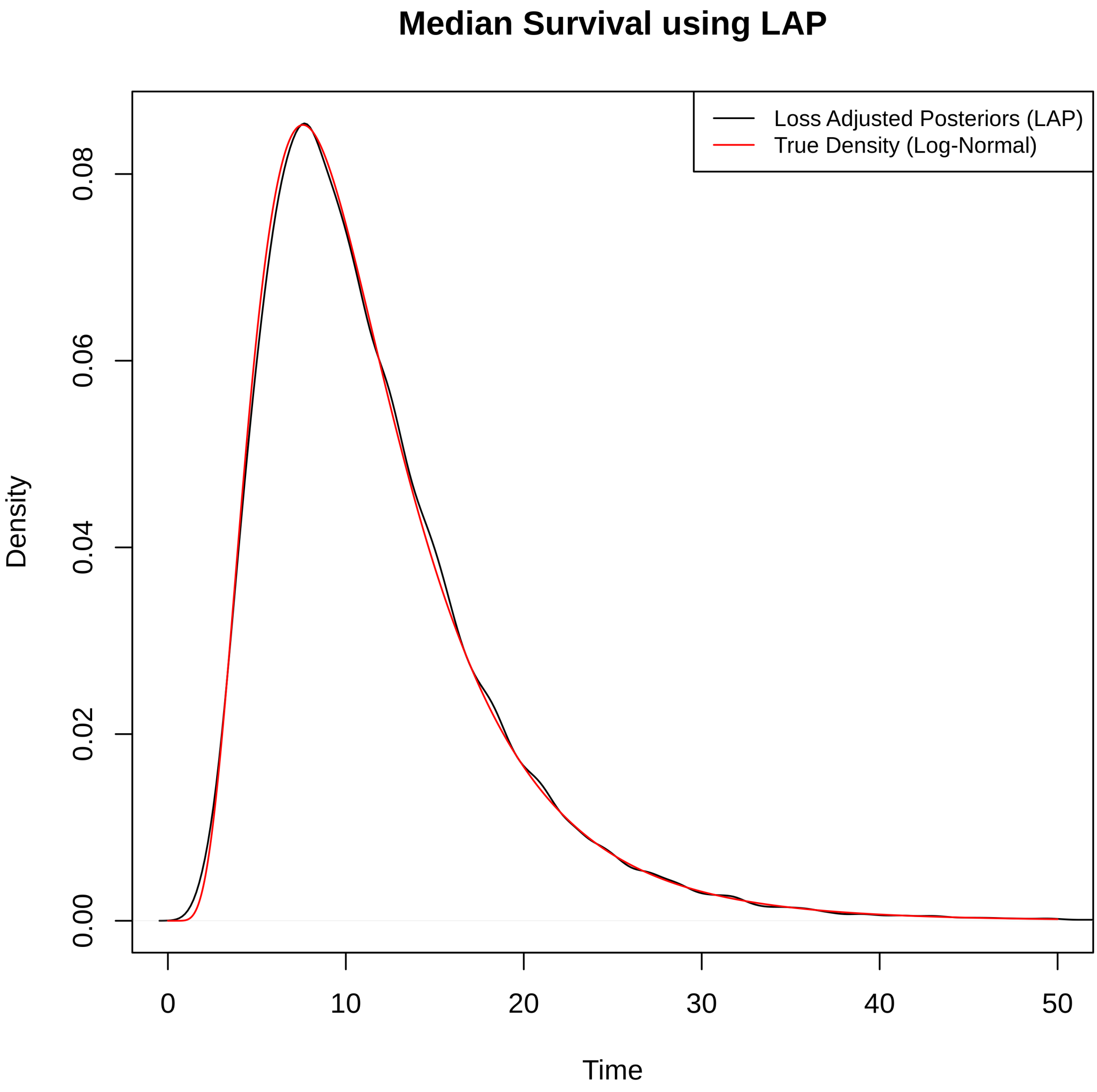


Figure 6.3: Posterior of Median Survival induced by loss function

Examples of proper and improper posteriors when including expert opinion

In this subsection we discuss a number of combinations of the prior and loss function which result in proper and improper posteriors. The loss functions all relate to expert beliefs about median survival for an exponential survival model with the λ or ψ parameterizations.

We first prove the propriety of the posterior distribution arising from the loss function in Equation 6.6 with a uniform prior on λ. The posterior density of λ in Equation 6.6 is $\pi(\lambda|\phi) \propto \mathcal{LN}(g(\lambda)|\mu_{\text{expert}}, \sigma_{\text{expert}}) \log(2)/\lambda^2$ across the interval $[a, b]$ with $a > 0$, $b < \infty$. In order to ensure this posterior is proper we need to prove the continuity of this function/density at each value of λ.

There are a number of properties relating to the continuity of functions and are detailed in (Fitzpatrick, 2009, Chapter 3). One of these is the product property, whereby if two functions f and g have a common domain D and both functions are continuous at x_0 then the product $f(x_0)g(x_0)$ is continuous at x_0. Additionally we are interested in composite functions, where function f has a domain D and g has a domain U such that $g(U)$ is contained within D. The composite function is denoted by $f(g(x))$ for all x in U. If $g(x)$ is continuous at x_0 and the function f is continuous at the point $g(x_0)$ the composite function is continuous at x_0.

Beginning with the part of the equation referring to the log-normal $\mathcal{LN}$ distribution, we see that it is a composite function. The domain of $g(\lambda)$ is the interval $[a, b]$ and it's image is the interval $[g(b), g(a)]$. The domain (or support) of the log-normal distribution is $(0, \infty)$, so the image of $g(\lambda)$ is contained within the domain of the log-normal distribution. The function $g(\lambda)$ is continuous within its domain and the log-normal distribution is continuous everywhere within its domain. Therefore, the composite function is continuous at every value of λ within $[a, b]$. The function $\log(2)/\lambda^2$ is also continuous for every value of λ within $[a, b]$ and, therefore, the product of these two functions is continuous.

Although we could have visualized this univariate function to verify its continuity at particular values of $\mu_{\text{expert}}, \sigma_{\text{expert}}$, the rationale provided above can be generalized to models with more than one parameter.

After establishing the propriety of the posterior including the expert opinion as a log-normal distribution we next consider the implications of updating improper priors with other loss functions. We begin by examining a situation where incorporating the loss function results in an improper posterior.

Assume an expert believes that median survival follows an exponential distribution with hazard parameter λ^*, distinct from λ which is the parameter for the survival model rather than the expert's belief. The mode of the exponential distribution is at 0, therefore, it results in non-zero density for λ as it approaches infinity. Assuming an improper uniform prior for λ (i.e. upper bound undefined), the posterior distribution is $\pi(\lambda|\phi) \propto \lambda^* \exp\{-\lambda^* \frac{\log(2)}{\lambda}\}$ with $\lim_{\lambda \to \infty} \pi(\lambda|\phi) = \lambda^*$. We can further extend this by stating that for the exponential model with the λ parameterization, any belief which puts a non-zero density on median survival equal to 0 (with an improper prior) will result in an improper posterior. It should be emphasized that while such beliefs are highly implausible, even opinions parameterized as valid probability distributions can under certain limited circumstances lead to improper posteriors.

While an improper prior combined with an exponential belief for median survival resulted in an improper posterior in the previous example, the use of an improper prior with other beliefs does not guarantee an improper posterior. For example, if the expert has a belief that median survival has a Cauchy distribution (truncated to be positive), the posterior is still proper even with an improper uniform prior. In this example we assume the exponential model has the ψ parameterization rather than the λ parameterization, as it allows us to evaluate the integral of the posterior density analytically. The posterior distribution of ψ is

$$\pi(\psi|\boldsymbol{\phi}) \propto \frac{1}{\pi\gamma\big[\big(\frac{\psi \log(2)-\mu}{\gamma}\big)^2+1\big]},$$

where μ, γ are location and scale parameters describing the density of the expert's belief. To evaluate the integral of this (unnormalized) posterior density with respect to ψ we first make the substitution $u = \frac{\psi \log(2)-\mu}{\gamma}$ for which $du = \frac{\log(2)}{\gamma} d\psi$. Expressed in terms of u and pulling constants outside the integral, the expression is now $\frac{1}{\pi \log(2)} \int \frac{1}{u^2+1} du$. This integral has an analytical expression and the indefinite integral in terms of u is $\frac{\arctan(u)}{\pi \log(2)} + C$ where C is the constant of integration. Undoing the substitution provides the final result which is

$$\frac{\arctan\big(\frac{\psi \log(2)-\mu}{\gamma}\big)}{\pi \log(2)} + C.$$

The limit of the arctangent of x when x is approaching infinity is equal to $\pi/2$, i.e. $\lim_{x\to\infty} \arctan(x) = \pi/2$. Therefore, the integral evaluates to a finite value even when the upper limit of integration not finite. The above integral was evaluated numerically for a number of values of μ, γ using the `integrate` function from the `stats` package (R Core Team, 2021) and gave identical results to the analytical expression of the integral.

A final example is how an improper prior for a parameter and an improper density describing an expert's belief can result in a proper posterior. We specify an improper uniform prior for λ over the interval $[a, \infty)$, where $a = 0.001$, and we also assume that the expert wishes to express complete ignorance regarding median survival times. This implies a uniform density on median survival over the interval $[0, \infty)$ which is improper. It should be noted that although the density on the expert's belief is improper, the prior on λ has restricted the possible values of the posterior median survival to be between $(0, \log(2)/a]$.

Although the expert has a belief corresponding to a uniform density for median survival, it implies a very non-uniform belief on λ. Because large values of median survival are plausible under this belief, the density of λ increases as it approaches zero (and is undefined at zero). By change of variable technique it is straightforward to show that the density for λ implied by a uniform belief on median survival is $f(\lambda) \propto \log(2)/\lambda^2$. As we have a uniform prior for λ, the posterior for λ is $\pi(\lambda|\boldsymbol{\phi}) \propto \log(2)/\lambda^2$ which is proper over the interval $[a, \infty)$.

ESS with an Exponential Likelihood

As with any elicitation exercise (either on the observable or parameter space) it is worth quantifying the ESS of the information. The ESS for an exponential likelihood under a gamma or inverse-gamma prior for λ or ψ parameters respectively is known through the conjugacy of the one-parameter exponential families (although the ESS will differ by 1 depending on the parameterization). Therefore it is most convenient to fit a gamma distribution to the percentiles of a posterior distribution for λ, using for example the `SHELF` package (Oakley, 2020), with the ESS informing the shape parameter.

6.5 Elicitation of parameters for a Multivariate Normal distribution

In this section we describe the incorporation of expert opinion for a complex statistical model to highlight the flexibility of the approach. We provide an approach which improves upon existing methods in that it requires fewer and less complex queries during the elicitation exercise.

6.5.1 Overview of previous approaches

The problem of including expert opinion within the multivariate normal sampling model has been explored using a natural conjugate prior (normal inverse-Wishart) and a non-conjugate prior (normal generalized inverse-Wishart) (Al-Awadhi and Garthwaite, 1998; Garthwaite and Al-Awadhi, 2001). The natural conjugate prior forces a dependence between the mean and the covariance, so Garthwaite and Al-Awadhi (2001) proposed assessment tasks that allow the expert to quantify separately assessments about each of these parameters.

In both approaches, assessment tasks include specifying conditional and unconditional percentiles and some assessments using based hypothetical observations. For example, the degrees of freedom parameters for a multivariate-t distribution, the prior predictive distribution for the multivariate normal, are assessed by considering the magnitude of difference between two random samples and assessing the median of this absolute deviation for each component Z_i. Then, experts are asked to suppose that two more observations are sampled from the population for which the magnitude of difference is calculated to be Z_i^*. These hypothetical values must not be what the expert was "expecting" (i.e., Z_i) and the expert must then assesses their conditional median of Z_i^*, with the ratio of these quantities used to calculate the degrees of freedom. The idea is that if the expert's conditional distribution Z_i^* changes by only a small amount relative to Z_i, then they have a strong belief about the spread of the multivariate distribution. As noted by Daneshkhah and Oakley (2010) “these are difficult assessments for the expert to make, as it is hard to judge how to change one’s beliefs in light of hypothetical data, particularly as this necessarily has to be done without writing down a prior distribution and applying Bayes’ theorem”. Furthermore, the method requires a substantial number of quantities to be elicited: for four-dimensional multivariate data, Daneshkhah and Oakley (2010) asked the expert to specify fifty quantities, a cognitively burdensome task.

6.5.2 Overview of proposed approach

Let $\mathbf{X}$ denote a k-dimensional random variable that has a $\mathcal{MVN}$ distribution, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Phi\ )$. The vector $\boldsymbol{\mu}$ is a k-dimensional parameter representing the mean of the distribution, while Φ is the $k \times k$ variance-covariance matrix. The covariance matrix is constructed from the correlation matrix Σ, which is pre- and post-multiplied by a diagonal matrix of the standard deviations $\mathbf{D} = \text{diag}(\sigma_1 \dots \sigma_k)$ so that $\Phi = \mathbf{D}\Sigma\mathbf{D}$.

The model priors are as follows:

$$\begin{aligned} \Sigma &\sim \text{LKJ}(\eta) \\ (\mu_i, \tau_i) &\sim \mathcal{NG}(\mu_0, \gamma, \alpha, \beta) \text{ for } i = 1, \dots, k, \end{aligned}$$

where the correlation matrix Σ is sampled from a Lewandowski-Kurowicka-Joe (LKJ) distribution (Lewandowski et al., 2009). This prior does not influence the variance components, unlike the Wishart distribution or its inverse parameterization. The mean μ_i and precision τ_i for each component X_i are modeled by a Normal-Gamma ($\mathcal{NG}$) distribution, with $\sigma_i = \sqrt{1/\tau_i}$ for $i = 1, \dots, k$.

We describe how to elicit the parameters of a multivariate normal distribution based on $k(k-1)/2 + 2k + 1$ elicitation steps. We first elicit the expert’s strength of belief, which we denote as n_e, by asking them to imagine the number of observations that their opinion represents, clarifying to them that each observation is a random sample of dimension k.

The subsequent steps of the approach can then be organised into three distinct sections: firstly, defining the hyperparameters for each of the k marginal normal distributions that make up the multivariate normal distribution; secondly, defining the loss function that encodes the expert’s belief about the pairwise correlation of the k elements; finally, adding a component to the loss function so that the prior on the correlation matrix does not attenuate the expert’s belief about the partial correlations.

Defining the hyperparameters for the marginal distributions

For each component $i = 1, \dots, k$, we model its mean μ_i and precision τ_i using an $\mathcal{NG}$ distribution. We can then write the conditional distribution as $\mu_i \mid \tau_i \sim \mathcal{N}(\mu_0, \gamma\tau)$, with the marginal distribution of the precision $\tau_i \sim \mathcal{G}(\alpha, \beta)$. The hyperparameter μ_0 represents the mean prior belief value of μ_i,[2] while γ denotes a scale factor which can be interpreted as the number of pseudo-observations or ESS (n_e) for each μ_i . Similarly the precision τ_i is estimated from 2α pseudo observations so that $\alpha = n_e/2$. The prior predictive distribution for an $\mathcal{NG}$ distribution is a non-standard, or scaled and shifted, Student's t distribution $\mathcal{S}t\left(\mu_0, \frac{\beta(\gamma+1)}{\alpha\gamma}, 2\alpha\right)$, where 2α denotes the degrees of freedom of the distribution (Bernardo and Smith, 2000). Because we have independence between the correlation matrix and other parameters, each component within the k-dimensional vector has the Student's-t distribution specified as above.

We proceed to identify the hyperparameters of the $\mathcal{NG}$ distribution as follows: for each component i in $1, \dots, k$, we elicit two percentiles of the prior predictive distribution, e.g. the median and upper quartile. With these parameters, along with n_e, we can find the parameters of the Student's t-distribution that optimally reflect these beliefs by minimizing the squared error.

Defining loss functions encoding beliefs about partial correlations

We now turn our attention towards the elicitation of the partial correlations ρ_{ij}, for $i, j = 1, \dots, k$. Fackler (1991) considers asking the experts for a concordance probability: $p_{ij} = \mathrm{P}(\theta_i > \mu_i, \theta_j > \mu_j)$ or $\mathrm{P}(\theta_i < \mu_i, \theta_j < \mu_j)$, i.e. the probability that both θ_i and θ_j are either above their expected values or below their expected values. For the bivariate case this probability is $p_{ij} = 0.5 + \sin^{-1}(\rho_{ij})$, with ρ_{ij} being the product moment correlation we require for the correlation matrix. Owing to the properties of the multivariate normal distribution, we can simply drop the elements not relating to i or j to obtain the bivariate distribution so that the formula holds for $k > 2$. See Kepner et al. (1989) for a simple derivation of the bivariate concordance formula which holds for the general multivariate normal distribution. While the prior probability of data arising from an $\mathcal{NG}$ distribution follows an $\mathcal{S}t$ distribution rather than a normal distribution, and no closed formulae exist for the concordance probability in this case, we have verified by simulation that their behaviour in this case is the same as that for the normal distribution.

We elicit the median concordance probability $\tilde{p}_{ij}$ for each combination (i, j) of the k variables, which is invariant to transformation. Assuming that the expert is as confident in

[2] Note that for $\alpha \leq 0.5$, the mean does not exist. In such cases, it is more accurate to refer to μ_0 as the median prior belief.

their beliefs about the concordance values as they are about the percentiles elicited above (though this is not a requirement), we use n_e to derive the uncertainty in the estimate of p_{ij}. We do this by first transforming $\tilde{p}_{ij}$ to $\tilde{\rho}_{ij}$, and then using Fisher's transformation $F(\rho) = \text{artanh}(\rho)$, the inverse hyperbolic tangent function, noting that this has a normal distribution with mean $\text{artanh}(\tilde{\rho}_{ij})$ and standard error $\psi = \frac{1}{\sqrt{n_e - 3}}$ (Fisher, 1915). Therefore the loss function includes terms $l(g(\rho_{ij}) \mid \phi) = -\log(\mathcal{N}(\text{artanh}(\rho_{ij})|\text{artanh}(\tilde{\rho}_{ij}), \psi))$ summed across all possible partial correlations.

Removing impact of prior for LKJ distribution

The marginal distribution of the partial correlation ρ_{ij}, modelled with an LKJ(η) prior is proportional to a $\mathcal{B}(\eta - 1 + k/2, \eta - 1 + k/2)$ beta distribution. This means that as k increases, the marginal prior probability for the partial correlations become non-uniform and more concentrated around 0. We can remove the impact of the prior on the correlation matrix by including the log density of the LKJ prior in the specification of the loss function, making it uniform over its support $[-1, 1]$. It is possible to obtain a uniform marginal by setting $\eta = (4 - k)/2$ so that we obtain a $\mathcal{B}(1, 1)$ distribution, however, clearly this is not possible for $k \geq 4$ as then $\eta \leq 0$. Figure 6.4 presents the marginal distribution for the LKJ prior with $\eta = 1, k = 4$ compared to the density of LKJ distribution included in the loss function.

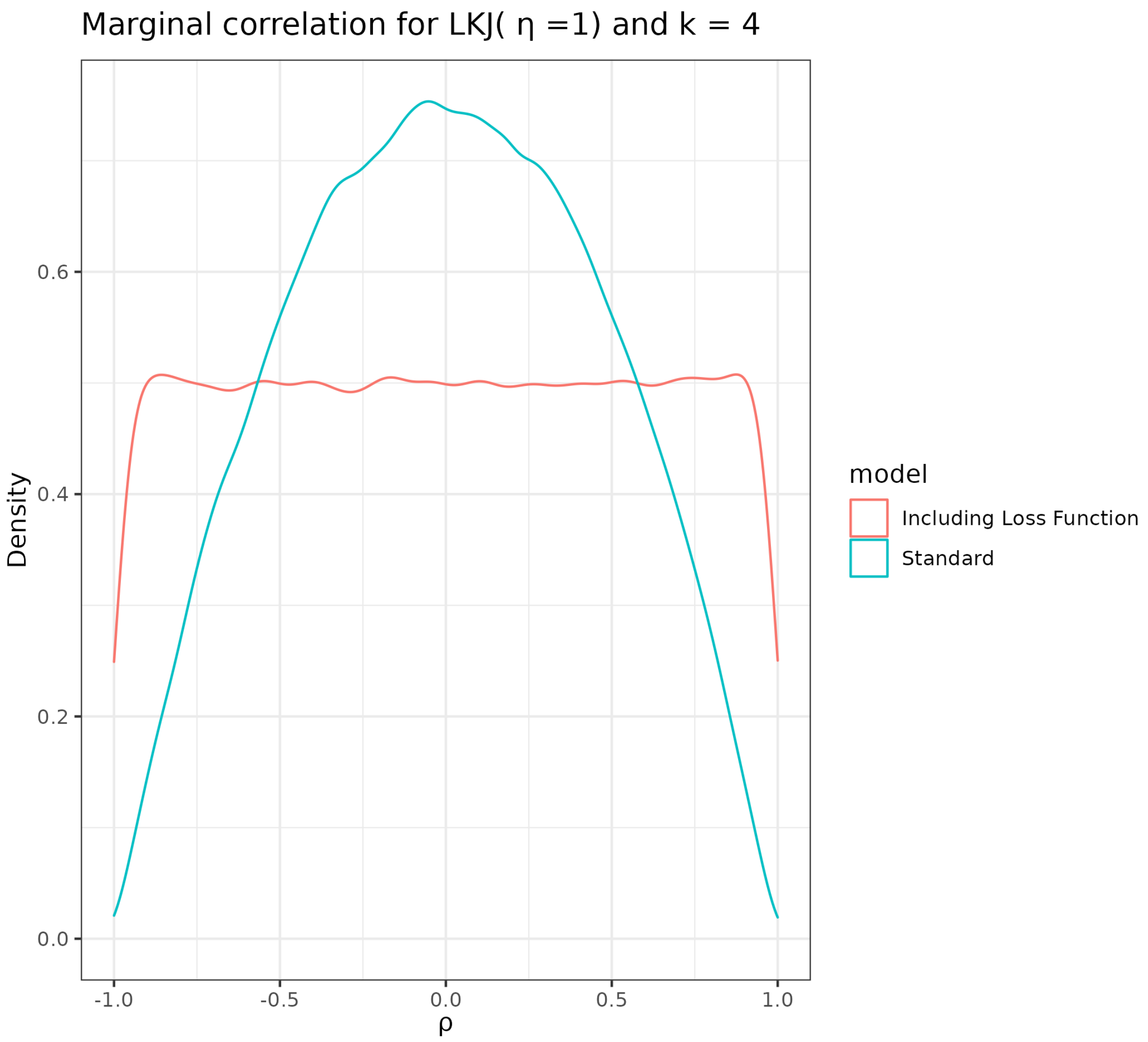


Figure 6.4: Comparison of densities when including of loss function vs. the standard specification of LKJ distribution

Summary

For the marginal distributions of each of the k elements we have defined hyperparameters based on elicited percentiles to match the expert's opinion. We include expert opinion on the correlation parameters through a loss function. Also included in the loss function is a term to negate the impact of the LKJ prior on the partial correlations.

Marginal Components of $\mathcal{MVN}$ distribution:

For each of the individual components $i = 1, \dots, k$ of the $\mathcal{MVN}$, find the hyperparameters of the associated $\mathcal{NG}$ distribution:

1. Elicit two candidate percentiles of the data distribution from the expert, e.g., median and upper quartile;
2. Ask the expert to describe their "confidence" in their belief in terms of an ESS,

noting that it can be different for each component i;

3. As the prior predictive distribution for a $\mathcal{NG}$ is a non-standard Student's-t, it is straightforward to find the hyperparameters of the $\mathcal{NG}$ distribution by numerical optimization using the information elicited above.

Correlation Parameters of $\mathcal{MVN}$ distribution:

1. For each pairwise combination (i, j) of the k elements, elicit the median concordance probability $\tilde{p}_{ij}$, the probability that both i and j will be above or below their respective medians;

2. Ask the expert to describe their "confidence" of their belief in terms of an ESS, noting that it can be different for each pairwise combination;

3. The median concordance probability $\tilde{p}_{ij}$ is transformed to the median partial correlation $\tilde{\rho}_{ij}$. Take the Fisher transformation, $\text{artanh}(\tilde{\rho}_{ij})$, of $\tilde{\rho}_{ij}$. This transformation of the partial correlation produces a variable whose distribution is approximately normally distributed, with mean $\text{artanh}(\tilde{\rho}_{ij})$ and a standard error $(\psi = \frac{1}{\sqrt{n_e - 3}})$, that is stable over different values of the partial correlation.

4. The loss function for each partial correlation is then

$$l(g(\rho_{ij}) \mid \phi) = -\log(\mathcal{N}(\text{artanh}(\rho_{ij})|\text{artanh}(\tilde{\rho}_{ij}), \psi)),$$

   and the total loss function is then

$$l(g(\boldsymbol{\rho}) \mid \phi) = \sum_{i<j}^{k} l(g(\rho_{ij}) \mid \phi) + \log(\text{LKJ}(\Sigma|\eta))$$

   where the inclusion of $\log(\text{LKJ}(\Sigma|\eta))$ ensures that $p(\rho_{ij})$ is uniform over the support $[-1, 1]$.

6.5.3 Expert Opinion applied to Multivariate Normal Model

We consider an example of the outlined method applied to an imagined elicitation exercise. We assume that the dimension of the multivariate normal data is $k = 4$, which necessitates 15 quantities to be elicited, in contrast to Garthwaite and Al-Awadhi (2001), who required > 50. The (hypothetical) expert has been asked to assess their effective sample size, which they believe to be $n_e = 10$. For each of the four marginal distributions, they specify the 0.5 and 0.75 percentiles as in Table 6.1.

Table 6.1: Percentiles supplied by the expert and associated hyperparameters.

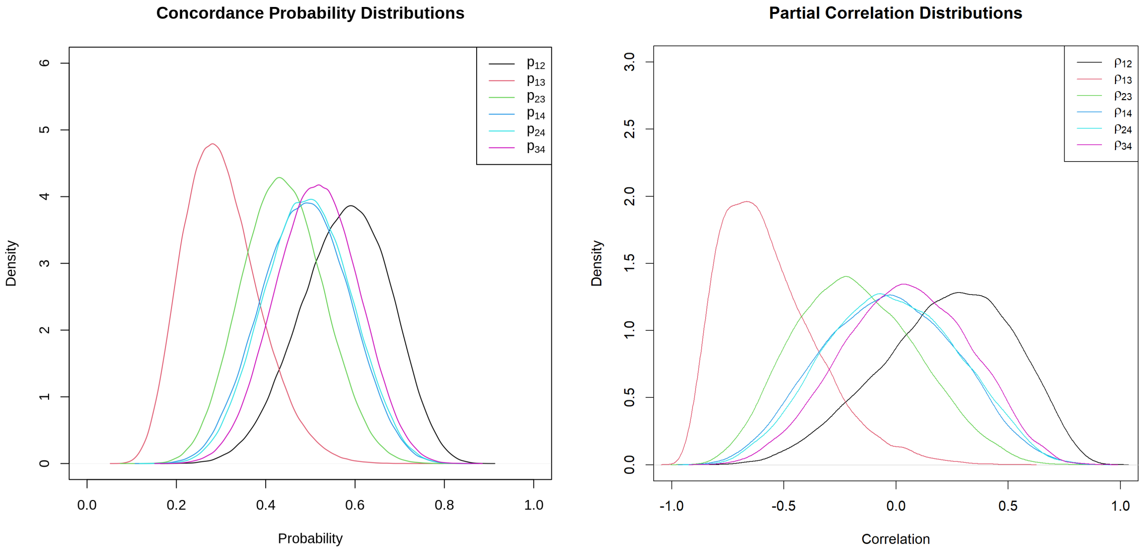
	Percentile		Hyperparameters	
k	0.5	0.75	μ_0	β
1	5.00	6.35	5	16.89
2	2.00	2.67	2	4.22
3	1.00	1.34	1	1.06
4	3.00	5.02	3	38

By definition $\gamma = n_e$ and $\alpha = n_e/2$.

For each of the six partial correlations, the expert provides their median concordance probabilities, as shown in Table 6.2. Also shown is the median posterior correlations estimated using the loss function. It is worth noting that the median concordance probability for $\tilde{p}_{13}$ from the model is higher than that specified by the expert. The reason for this is that for any one partial correlation, conditional on the other partial correlations, the remaining partial correlation is restricted to be within a certain interval so that the correlation matrix is semi-positive definite. We suggest that once the median concordance probabilities are elicited, that the intervals for each concordance probability which produce a positive definite correlation matrix, conditional on the other concordance probabilities, are presented to the expert. In situations where there is substantial density outside the originally specified interval, the expert should asked to reassess this particular value to ensure coherency. Furthermore, the expert can be shown the distributions of the concordance probabilities and the correlation parameters induced by incorporating the loss function and confirm that it is a reasonable representation of their beliefs. These plots are shown in Figure 6.5.

Table 6.2: Median concordance probabilities supplied by the expert and those generated by the model.

	Expert's concordance probability	concordance probability induced from model
$\tilde{p_{12}}$	0.60	0.58
$\tilde{p_{13}}$	0.25	0.30
$\tilde{p_{23}}$	0.40	0.44
$\tilde{p_{14}}$	0.50	0.49
$\tilde{p_{24}}$	0.50	0.49
$\tilde{p_{34}}$	0.50	0.51



(a) Concordance Probability

(b) Partial Correlation

Figure 6.5: Concordance and Correlation induced by loss function

6.6 Repeated measurements regression model

In our final example we consider a repeated measures regression problem. To our knowledge, this class of problem has yet to be analysed using expert opinion on the observable space, but the outlined approach can be accomplished by adding a few lines of code to popular existing software.

To motivate our approach, we consider data presented by Littell (1990), in which the effect of three different exercise programs, denoted as CONT, RI and WI, on participant's strength is assessed over seven different timepoints. In the original analysis the population effects (often called fixed-effects) were modelled with time as a quadratic function, with various covariance structures under consideration. We consider a situation in which the expert was asked for their belief about the *expected* change from baseline for the WI programme.

We let y_{ij} denote the measurement of the ith person at the jth timepoint. To account for the repeated measures we define individual effects (also known as random effects) for the intercept, slope, and quadratic effect of time, along with population level effects for the programme term and an interaction term between linear and quadratic time and the programme. The full model has the essential form:

$$\begin{aligned} y_{ij}|\mathbf{b_i} &= (B_0 + b_{0i}) + (B_1 + b_{1i})t_j + (B_2 + b_{2i})t_j^2 \\ &+ B_3 P_{WI} + B_4 P_{WI} t_j + B_5 P_{WI} t_j^2 + \cdots + \epsilon_{ij}. \end{aligned}$$

The population level intercept is denoted by B_0, while B_1 and B_2 denote the population level coefficients for linear and quadratic time. The other population level terms refer to programme related time effects:

$$P_{WI} = \begin{cases} 1 & \text{treatment group is WI} \\ 0 & \text{otherwise} \end{cases}$$

and for space we have suppressed the population effects for the P_{RI} relative to baseline group (CONT).

The individual level effects $\mathbf{b_i}$ for the intercept and slope of linear and quadratic time are modelled as a multivariate normal distribution with zero mean and covariance matrix

$\mathbf{b_i} = (b_{0i}, b_{1i}, b_{2i})^\intercal \sim \mathcal{MVN}(0, \Phi)$.

6.6.1 Expert Opinion

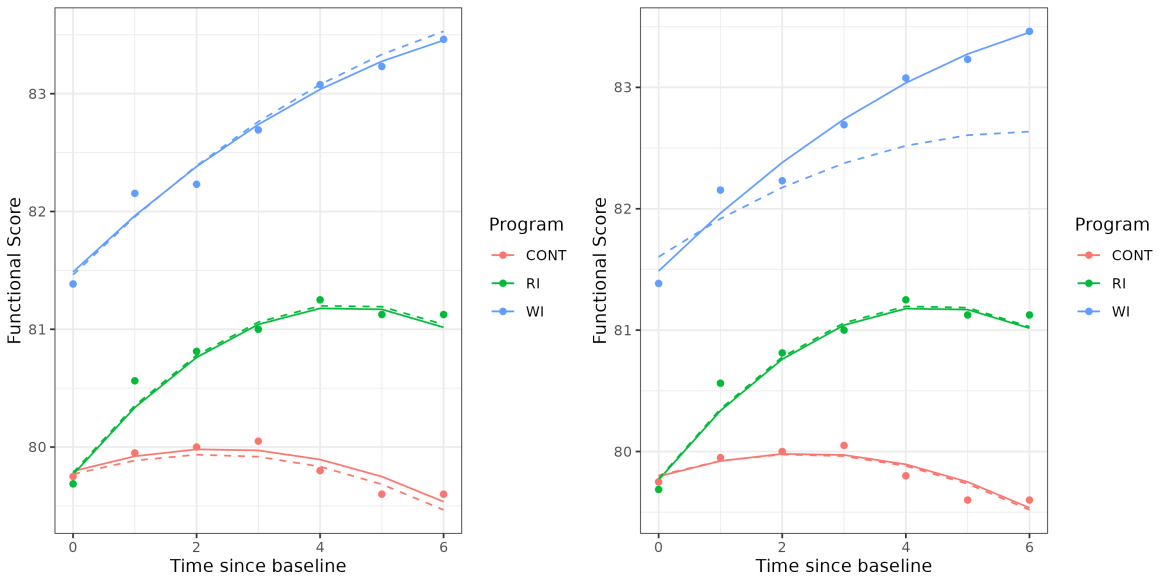
We suppose that expert opinion for the WI group is available whereby the expert believes that the mean change in strength from baseline to the final timepoint for this group can be represented as a normal distribution, with mean $\mu_{expert} = 2.5$ and standard deviation $\sigma_{expert} = 1.5$. We also consider a scenario analysis where the expert opinion is much stronger, with $\mu_{expert} = 0.5$ and $\sigma_{expert} = 0.5$. Because of the quadratic term, it is not possible to place a prior on the population level coefficients to induce these opinions. However, it is straightforward to do so using a loss function.

The difference in expected strength between final and first timepoints for the WI group is $\zeta = B_4 P_{WI} t_j + B_5 P_{WI} t_j^2$. Our loss function then amounts to

$$l(\xi, \mu_{\text{expert}}, \sigma_{\text{expert}}) = \frac{1}{2}\left(\frac{\zeta - \mu_{\text{expert}}}{\sigma_{\text{expert}}}\right)^2.$$

We specify time (linear and quadratic components) as orthogonal polynomial contrasts to aid estimation. Stan code for this model was generated using the `brms` package (Bürkner, 2017) which was then modified to include the loss function. Vague (but proper) priors were chosen for all model parameters and a LKJ prior with $\eta = 1$ was chosen for the correlation matrix. The impact of the expert's opinions are illustrated in Figures 6.6 & 6.7. In the first scenario, the expert's opinion has very marginally changed the fixed effects regression line while for the scenario with $\mu_{expert} = 0.5$ and $\sigma_{expert} = 0.5$ the impact is substantial. The changes in the 95% credible intervals are also apparent in the posterior distributions for the expected change from baseline.

We can get a heuristic for the ESS of the expert's opinion by comparing the standard deviation of the posterior distribution for the change from baseline without expert opinion



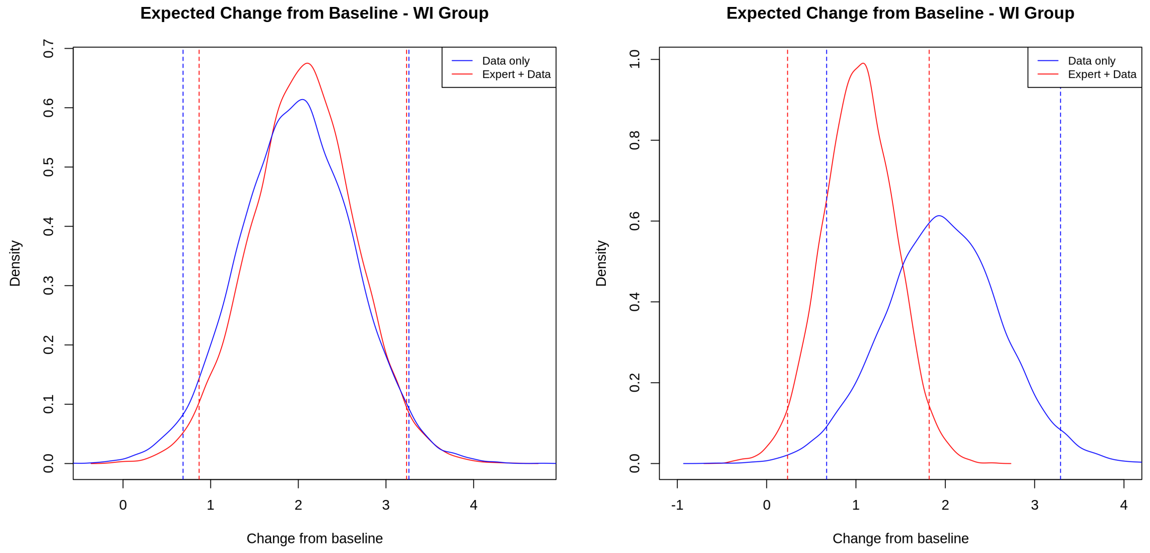
(a) Expert Opinion:$\mu_{\text{expert}} = 2.5, \sigma_{\text{expert}} = 1.5$ (b) Expert Opinion:$\mu_{\text{expert}} = 0.5, \sigma_{\text{expert}} = 0.5$

Figure 6.6: Comparison fixed effect model estimates with and without expert opinion (dashed and full line respectively).

(0.67) to that of the standard deviation obtained using the expert opinion (1.5). Considering that the value of 0.67 was generated from 13 participants in the WI group, we get the relation $\sigma = 0.67\sqrt{n_{\text{data}}} \approx 1.5\sqrt{n_{\text{expert}}}$ so that $n_{\text{expert}} \approx 2.5$. This value seems plausible considering the modest impact of including the data. In the scenario in which the expert opinion has standard deviation of 0.5, a similar calculation produces $n_{\text{expert}} \approx 23$. We note that, similar to Morita et al. (2008) and Neuenschwander et al. (2020), this calculation ignores the potential for the expert opinion to be in conflict with the observed data as is potentially the case in the second scenario. However, it is still useful as a way to highlight the large change in the strength of the respective expert opinions.

6.7 Discussion

In this chapter we describe a general approach to including expert opinions on observable quantities within statistical models. The theoretical justification for the approach is based on Bissiri et al. (2016), and expands upon this framework in a theoretical sense and by providing several practical examples. Bissiri et al. (2016) briefly describe an example in which an expert declares that a parameter θ is close to 0 with a quadratic loss function. The core idea is that the expert declares that a function of a parameter (which relates to the observable quantity) can be described in some manner such as a probability distribution. Additionally, this chapter describes the potential for the prior on the parameters to attenuate the information provided by the expert and how to eliminate its impact. As shown in Section 6.5.3 the specification of a loss function may be of interest



(a) Expert Opinion:$\mu_{expert} = 2.5, \sigma_{expert} = 1.5$ (b) Expert Opinion:$\mu_{expert} = 0.5, \sigma_{expert} = 0.5$

Figure 6.7: Comparison of mean change from baseline estimates with and without expert opinion (dashed lines refer to the 95% credible intervals).

in standard data analysis exercises, such as when attempting to model multivariate data in which the correlations are close to ± 1, as with higher dimension of the data, the LKJ distribution will place more prior probability on partial correlations close to zero.

In Section 6.4.3, we outline general assumptions for ensuring the propriety of the Loss Adjusted Posterior. Specifically, assuming a proper prior for the model parameters and that the expert's opinion results in a continuous density across the parameter space, we expect the Loss Adjusted Posteriors to be proper. Similar to the standard application of Bayes' theorem, where a proper prior multiplied by a likelihood yields a proper posterior, we posit that combining a proper prior with expert opinion (expressed through a loss function) will similarly lead to a proper posterior. However, we acknowledge that this assumption remains unproven. When updating improper priors with data, it is important to verify the propriety of the resulting posterior. The same caution applies when incorporating expert opinions. As in the exponential example in Section 6.4.3, care must be taken to ensure that non-zero density is not placed at infinite values of the parameter space. To mitigate this risk, we recommend avoiding improper priors altogether. Fortunately, opinions related to observable outcomes naturally constrain the parameter space, allowing us to include vague yet proper priors. For instance, our uniform parameter specification for λ from 0.001 to 10 accommodates median survival values between 0.069 and 693, which should encompass all plausible values.

The approach considered in this chapter requires experts to assign subjective probability to deterministic functions of the parameters such as expected values or percentiles and therefore, is in the same category as work by Bedrick et al. (1996) and Hosack et al. (2017), albeit acknowledging that the focus is not on the prior but a loss function. Within the LAP framework it is not possible to specify beliefs about the prior predictive density as by definition the parameter of interest has been integrated out of the expression (and we cannot specify a deterministic relationship between the data x and the parameter(s) which generated that data). This however does not prevent using the prior predictive approach to define priors which encode expert opinion for some of the parameters with information on other parameters included through the loss function (as in Section 6.5.3).

The approach described in this chapter addresses many of the key difficulties with respect to expert knowledge elicitation described by Mikkola et al. (2023). They cite practical reasons, such as many of the approaches still being too difficult for non-statistical experts to use, and the lack of good open source software that integrates well with the current probabilistic programming tools used for other parts of the modelling workflow. It is natural to assume that eliciting expert's opinions on observable quantities is easier than elicitation on the parameter space, however, it can still be cognitively burdensome. As described in Section 6.5, our approach to elicitation of the multivariate normal distribution requires many fewer questions and avoids elicitation about conditional distributions and hypothetical data which are cognitively more challenging. Another key issue that we believe this method addresses is that it integrates well with the current probabilistic programming tools used for other parts of the modelling workflow. Stan and other Bayesian programs are the default tools for many statisticians and the ease which this approach can be integrated with these tools is highlighted by the pseudo-code in Section 6.4.3. Only a few lines of code were required to update the existing code generated by the brms package for the example in Section 6.6.

Using this approach, it is typically straightforward to include expert opinion on observable quantities, however, it always worth checking how well the posterior distribution for the expert's opinion (just with the loss function and without data) approximates the opinion elicited from the expert. As noted earlier, a uniform prior on the parameters does not necessitate a uniform density on a function of those parameters such as the observable quantity. In many practical situations the choice of vague priors will have a relatively modest impact on the density implied by the expert, which will diminish the more informative the expert's opinion is. Furthermore, in the context of survival analysis, Cooney and White (2023a) found that posterior distributions estimated using data and loss functions tended to be very similar even with different types of relatively non-informative priors[3]. In the situation that the impact is non-trivial, it is often possible to express the observable quantity upon which expert opinion was sought as a parameter,

[3] They also found very good agreement between the Bayesian method and a frequentist approach motivated as a penalized likelihood.

assign a uniform prior to it and express one of the model parameters as a deterministic function of the observable quantity (and the other parameters as required) as described in Section 6.4.3. Even if it is not possible to express this relationship analytically, it is usually possible to solve this numerically, albeit with a large increase in computational burden.

Finally, although the incorporation of expert opinion with statistical models using this approach is more straightforward, the robustness of the inferences generated from them will rely upon the quality of the information elicited from the experts. O'Hagan et al. (2006) provides an in-depth treatment of expert elicitation and it is evident that considerable effort, sometimes taking the form of a workshop, is required to obtain methodologically appropriate opinions. Although important to quantify (at least approximately), the informativeness of any type of elicited information included in a statistical model, it is especially important in situations where the expert is providing opinion on the expected value of the response and their uncertainty around that estimate (e.g. Section 6.6) or elicitations based on prior predictive quantiles. Experts may not understand that uncertainty does not scale linearly with the sample size, as illustrated in the extreme case in Figure 6.1. Therefore in each of our examples we try to produce an approximate estimate of ESS (or as in the case of Section 6.5.3 specify it explicitly) so that the expert can be given the opportunity to revise their estimates at the elicitation stage rather than requiring post-hoc adjustments or reassessments.

Chapter 7 will apply the framework we have developed here to the full range of parametric survival models, exploring approaches to pooling multiple experts' opinions and quantifying their relative strength in terms of data observations.

7 Utilizing Expert Opinion to inform Extrapolation of Survival Models

As noted Chapter 2 parametric survival models extrapolate the observed survival data to make long-term survival projections that are crucial to cost-effectiveness decision making. Differences in long-term predictions can be particularly pronounced when a high proportion of the survival times are censored and may produce clinically implausible survival estimates (Davies et al., 2013). When expert clinical opinion is available, it is important to use this information in the modelling process. Often these opinions are not integrated in a formal way, with survival models typically estimated using maximum likelihood, i.e., based on the data alone, before choosing the parametric model for which expected survival appears to be compatible with the expert opinion. This approach has a number of weaknesses. Primarily, it is difficult to identify the most appropriate model if several models appear consistent with the expert opinion. In the opposite scenario, when none of the models meet the expert's criteria, the best choice of model is again unclear. It would be preferable to have a measure of statistical fit which takes account of the degree of agreement with the expert opinion as well as the observed data, rather than making a decision based solely on whether the predicted quantity from the model is within the expert's plausible range. In this chapter we consider how long-term survival estimates provided by clinical expert opinion can be directly incorporated into the model estimation procedure. We do so by adopting a framework in which the expert opinion that has been elicited on the observable data space is used to modify the density of the parameter space. Our approach is compatible with the SHELF elicitation framework (Oakley, 2020), including when multiple expert opinions are available, and can be applied to many parametric models, from the exponential distribution, which assumes a constant hazard, to spline models that can accommodate bathtub type hazards. This approach generalises previous work by Cope et al. (2019), in that we consider the parametric survival models commonly used in decision making, evaluate model fit based on goodness of fit to both data and expert opinion and do not restrict expert opinion to be represented by a single normal distribution. The rest of the chapter is organized as follows. We provide a review of some methods which incorporate expert opinion into parametric survival models. Subsequently, we introduce the proposed statistical method, and discuss considerations when aggregating the opinions of multiple experts. We then present an application of the method whereby the survival times of pediatric acute lymphoblastic leukaemia (pALL) patients treated with tisagenlecleucel are integrated with expert opinions about survival at various timepoints (Grupp et al., 2016, 2018). We conclude the chapter with a discussion of its key ideas along with a summary of the challenges involved in eliciting expert opinion. In Appendix D.1 we validate our approach with a previously published example whereby expert opinion on median life was incorporated into the survival function. In Appendix D.2 we describe some technical details regarding the estimation of certain parametric models. In Appendix D.4 we perform a simulation study to assess the effect of priors on posterior survival extrapolations when including expert opinion and in Appendix D.5 we perform a simulation study to assess the impact of bias in expert opinion on extrapolated survival. All methods outlined in this chapter are available for use as an R package called `expertsurv` (Cooney and White, 2023b) with further details presented in Chapter 8.

7.1 Previous Literature

Much of the initial work on this topic is from reliability analysis, incorporating expert opinion about the median survival into Weibull models, with the median survival distributed as a function of chi-squared distribution or a normal distribution (Campodonico and Singpurwalla, 1993; Singpurewalla and Song, 1988). One disadvantage of both approaches is that the experts are also required to think about the mean and variance of the shape parameter of the Weibull distribution (i.e. parameter space), which is much more difficult than eliciting information in the observable data space. Other work estimated the Weibull model based on expert opinion from either the mean, mode and quantiles of survival time and a hyperparameter representing the effective sample size of the opinion, avoiding the need to elicit expert opinion on the parameter space (Bousquet, 2006). Another approach uses hyperparameters to incorporate expert opinion for survival models for which conjugate priors or priors with the same form as the likelihood exist (exponential, gamma, and Weibull). For two-parameter models, however, this approach requires the assumption that one of the parameters is already known (Coolen, 1996). The approach generates informative priors by calculating their hyperparameters using sufficient statistics such as (but not limited to) the number of events, censored observations and the sum of event times. Survival models with covariates can also incorporate expert opinion, whereby the expert contributes a distribution conditional on the values of the covariates at a design point. As described in Chapter 6 for a class of priors referred to as data augmentation priors (DAP) in which the prior has the same form as the likelihood, expert opinion was incorporated at different levels of a covariate for exponential and log-normal examples, again through deriving hyperparameters (Bedrick et al., 1996; Johnson, 1996). In the context of health technology assessment, Ouwens (2018) incorporated expert opinion about survival probabilities at a particular timepoint for one and two parameter models by re-expressing one of the parameters as a function of the survival probability at the chosen timepoint and the other parameter (if applicable). This approach considers a broader family of parametric models than those previously described. The approach samples both a survival probability from the expert's prior distribution and the second parameter from its (non-informative) prior and uses these to calculate the first parameter. A similar hierarchical Bayesian approach (although from the field of ecology) considered the Weibull model with expert opinion elicited on mean survival at different covariate levels for multiple experts (Wongnak et al., 2022). Willigers, Bart and Ouwens, Mario and Briggs, Andrew and Heerspink, Hiddo and Pollock, Carol and Pecoits-Filho, Roberto and Tangri, Navdeep and Kovesdy, Csaba and Wheeler, David and Garcia-Sanchez, Juan Jose (2023) extend the approach of Ouwens (2018) allowing for two timepoints and in situations when model parameters which cannot be analytically expressed in terms survival probabilities optimizer is used to obtain the parameters. Cope et al. (2019) introduced a method to incorporate expert information regarding survival probabilities when it has been provided at multiple timepoints. In the approach of Cope et al. (2019) a Bayesian approach is used to fit a hazard function to the observed data and the hazards implied by the long-term survival beliefs of the expert. Weibull, Gompertz, 1st and 2nd order fractional polynomials can be fit with this approach using the JAGS statistical program(Plummer, 2003), however, it is not clear if the expert opinion modifies the model parameters or if it is solely the hazards implied by the expert's survival beliefs which are used to extrapolate the survival beyond the observed data. Ayers et al. (2022) implement expert opinion assuming that the expert's belief about survival at a particular timepoint is normally distributed (truncated at zero). This approach treats the elicited mean value of the expert's belief for survival as a datapoint. This datapoint is assumed to be generated from a truncated normal distribution with mean equal to the predicted survival based on the model parameters and standard deviation based on the normal distribution elicited from the expert. It can be shown that this statement will correctly incorporate the expert's belief only for symmetrical non-truncated distributions.

Che et al. (2023) provide a method to incorporate an opinion by simulating a dataset of survival times with the expected survival as per the expert's belief with the sample size providing the strength of expert opinion. If the expert believes 20% will survive beyond 10 years and their opinion is equivalent to 100 observations, then 80 observations from a $\mathcal{U}(0, 10)$ (uniform distribution) can be simulated and remaining 20 observations censored.

This approach, however, implies a functional form for the survival times (i.e. uniform) which may not be consistent with the observed survival data. Recently, Jackson (2023) estimated spline models in which expert opinion on conditional survival parameterized as a $\mathcal{B}$ (beta distribution) can be incorporated at multiple timepoints.

7.2 Survival Analysis with Expert Information

Using the notation from Section 2.3, consider an exponential distribution, with associated hazard $h(t) = \theta$ and survival function $S(t) = \exp\{-\theta t\}$. The likelihood of an exponential model is then $L(\theta|D) = \theta^{\sum_{i=1}^{n} \nu_i} \exp\left\{-\theta \sum t_i\right\}$. If the prior distribution for θ has been specified as $\mathcal{G}(\alpha, \beta)$, i.e., a Gamma density with parameters α and β, then the posterior distribution is available in closed form as Gamma distribution $\mathcal{G}(\alpha + \sum_{i=1}^{n} \nu_i, \beta + \sum_{i=1}^{n} t_i)$. While in this case the posterior distribution is tractable, Bayesian inference for other distributions is more challenging and relies on modern computational methods for inference. Even in this case, tractable inference requires specification of the prior in a specific framework that will not be intuitive to a non-specialist. See Table D.6 in Appendix D.6 for a full list of the survival models under consideration in our analysis.

7.2.1 Integrating Expert Opinion with Trial Data

Consider the situation where an expert has an opinion about the survival probability at potentially multiple times $\mathbf{t}^* = t_1^*, \ldots, t_k^*$. As in the Chapter 6 we propose to incorporate this information into the analysis by expressing the elicited quantity in terms of the parameters $\boldsymbol{\phi}$ which will contribute a "loss" or penalization based on the discrepancy with the elicited opinion. The parameters $\boldsymbol{\phi}$ will typically be parameters of a specific probability distribution describing the expert's opinion about the survival at each of the times $\mathbf{t}^*$. The parameters will be estimated based on their fidelity to the data and expert opinion, with the relative strength determined by the number of observations and precision of the elicited belief.

In the most general situation in which we have k timepoints at which we wish to include expert opinion, we let ϕ_i represent parameters associated with the timepoint i and $l(g(\theta), t_i^*, \boldsymbol{\phi}_\mathbf{i})$ a loss function encoding expert opinion at timepoint i. The posterior distribution of the model parameters including expert opinion is: $\pi(\theta|\boldsymbol{\phi}) \propto \exp\{-\sum_{i=1}^{k} l(g(\theta), t_i^*, \boldsymbol{\phi}_\mathbf{i})\} p(\theta)$.

As in Chapter 6, $g(\theta)$ is a function of the model parameters θ and the parameters governing the experts' opinion $\boldsymbol{\phi}_i$ at a particular timepoint t_i^*. To fix this idea, consider an exponential model being fit to data, with a normal distribution with mean μ_i and standard deviation σ_i (ϕ_i) describing the expert's belief about survival at a particular timepoint t_i^*, so that $S(t_i^*) \sim \mathcal{N}(\mu_i, \sigma_i)$ and $g(\theta) = S(t_i^*)$. The posterior density is then proportional to:

$$\pi(\theta, \boldsymbol{\mu}_{1\ldots k}, \boldsymbol{\sigma}_{1\ldots k}) \propto \exp\big\{-\sum_{i=1}^{k} l(g(\theta), t_i^*, \mu_i, \sigma_i)\big\}p(\theta),$$

and the loss function for the expert opinion at the particular timepoint is

$$l(\theta, t_i^*, \mu_i, \sigma_i) \propto \frac{1}{2}\bigg(\frac{\exp(-\theta t^*) - \mu_i}{\sigma_i}\bigg)^2.$$

The posterior of the expert opinion with data is denoted by

$$\pi(\theta|D, \boldsymbol{\mu}_{1,\ldots,k}, \boldsymbol{\sigma}_{1,\ldots,k}) \propto \exp\big\{-(\sum_{i=1}^{k} l(g(\theta), t_i^*, \mu_i, \sigma_i) + l(\theta|D))\big\}p(\theta),$$

with the posterior includes the loss function from the (negative) data log-likelihood and the (negative) log-density. Considering the survival at one timepoint (suppressing the subscript i) with t^* based on the μ and σ parameters representing the expert opinion for that timepoint. Finally $p(\theta)$ denotes a weakly informative prior for θ. For an exponential model the survival at the elicited timepoint is $S(t^*) = \exp\{-\theta t^*\}$ so that

$$l(\theta, \mu_{\text{expert}}, \sigma_{\text{expert}}, D) \propto -\Big\{-\frac{1}{2}\bigg(\frac{\exp(-\theta t^*) - \mu_{\text{expert}}}{\sigma_{\text{expert}}}\bigg)^2 + \sum_{i=1}^{n} \nu_i \log\theta + -\theta\sum t_i\Big\}.$$

While the resultant posterior does not have a closed form, this is not of practical importance when using modern computational Bayesian methods. More generally, the advantage of this approach is that it can be applied to a wide family of survival models, including those with 3 or more parameters. It is also straightforward to represent the elicited opinion as other probability distributions e.g. beta distribution, as well as incorporate additional timepoints. As detailed extensively in Chapter 6, theoretical justification for this approach is provided by Bissiri et al. (2016) who show that a valid and coherent update of a prior belief distribution to a posterior can be made for parameters which are connected to observations through a loss function rather than the traditional likelihood function, which is recovered as a special case. Although we have presented this method in a Bayesian framework, it can also be motivated from a frequentist perspective as an example of a penalized likelihood method (Cole et al., 2013). In this framework, we impose additional constraints on the parameter space by modifying the likelihood so that it is a function of the observed data and a further penalty term that will pull or shrink the final estimates away from the maximum likelihood estimates

(MLEs), towards values of the parameters which are more compatible with the elicited predicted survival at the timepoint(s) $\mathbf{t}^*$. Model estimation with this approach can be achieved using standard optimisation techniques (Nocedal and Wright, 2006).

When considering the analysis in a Bayesian framework it is worth discussing the potential effect of the prior on the parameters $p(\theta)$ and the loss function which encodes the expert's opinion $l(\theta, \mathbf{t}^*, \phi)$. As described in Chapter 6 the information encoded in the loss function is distinct from the prior, however, it is possible that a particularly informative prior on the model parameters could also imply a density for survival at the designated timepoints which conflicts with the expert's opinion. We conducted an extensive simulation study, with various sample sizes and expert opinions (in terms of location and spread of the beliefs) for the parametric models we have implemented. We compared results of the models fit using uniform priors for all parameters to relatively vague normal and gamma priors which were more informative than those typically used in Bayesian analysis. Across the scenarios specified in the simulation study, the posterior distribution of the survival functions were effectively identical. Additionally, we compared the results versus those obtained through penalized maximum likelihood estimation so that we could compare with a method which does not include a prior. We again obtained very similar results. From this we can conclude that reasonably non-informative priors for the parameters should not conflict with information provided by an expert. Further details regarding the simulation studies are presented in Appendix D.4.

7.2.2 Incorporating Multiple Expert Opinions

In some situations the opinions of multiple experts are available and in general groups tend to perform better than the average individual in elicitation exercises (Clemen and Winkler, 1999). Although it is sometimes reasonable to provide a decision maker with the elicited expert probability distributions separately, the range of which can be studied using sensitivity analysis, it is often necessary to combine the distributions together into a single analysis. In many cases, for example, a single distribution is needed for input into a larger model; and that model has other inputs with structural uncertainties, so that a full sensitivity analysis may not be feasible. This is particularly true when a specific survival model is used as an input for a cost-effectiveness model, in which case decision makers are typically making choices with respect to the parametric model in question, in addition to other structural assumptions. Considering this model choice appropriately for each expert can be burdensome and inflate the number of scenarios presented to the decision maker.

Combination, or aggregation, procedures are often dichotomized into mathematical and behavioural approaches, although in practice aggregation might involve some aspects of each (O'Hagan et al., 2006). One such technique is opinion pooling, where a *consensus distribution* for $p(\theta)$ is obtained as some function of the distributions $p_1(\theta), \dots, p_m(\theta)$ elicited from each of the m individual experts. Behavioural aggregation approaches attempt to generate agreement among the experts by having them interact in some way. This interaction may be face-to-face or may involve exchanges of information without direct contact or have an impartial observer to facilitate discussion (e.g. SHELF protocol). In either case the consensus distribution is then used as the prior for the analysis. For the purpose of this chapter we will focus on opinion pooling (as to conduct behavioural aggregation would require us to conduct an elicitation exercise with experts) noting that these methods are simpler to implement than behavioural approaches, although the distributions that result need not represent the opinions of any one person, let alone a consensus opinion from the group of experts. Additionally, these methods have "coherency" issues, as highlighted below.

We first consider the logarithmic opinion pool, which is obtained by taking a weighted *geometric* mean of the distributions,

$$p(\theta) \propto \prod_{j=1}^{m} p_j(\theta)^{w_j},$$

with weights specified such that $\sum_{j=1}^{m} w_j = 1$. When the decision maker is equally confident in the abilities of all experts, it is common to choose $w_j = 1/m$ for all j. The advantage of this approach is that it is *externally Bayesian*. When new data are obtained, one could either update each expert's distribution individually and then combine the resulting posterior distributions using logarithmic pooling, or first combine the expert's distributions and then update the consensus distribution. These will result in the same posterior distributions.

Continuing our example and assuming an exponential distribution with constant hazard, if m experts have expressed their prior beliefs about θ as Gamma priors $\mathcal{G}(\alpha_j, \beta_j)$, $j = 1, \dots, m$ the pooled prior is also a Gamma distribution, $\mathcal{G}(\sum_{j=1}^{k} w_j\alpha_j, \sum_{j=1}^{m} w_j\beta_j)$, and the resulting posterior distribution is then $\mathcal{G}(\sum_{j=1}^{m} w_j\alpha_j + \sum_{i=1}^{n} \nu_i, \sum_{j=1}^{m} w_j\beta_j + \sum_{i=1}^{n} t_i)$. If we were to compute the posterior distribution using each expert prior separately, and then compute the logarithmic opinion pool post-hoc, it is evident that the same posterior distribution would be obtained.

Logarithmic pooling, however, does not satisfy the *marginalisation* property. Suppose each expert is asked about mutually exclusive events, A and B. If C is the event A or B, then coherency demands that $Pr(C) = Pr(A) + Pr(B)$. There are two ways to obtain a pooled probability for C. We can compute the probability by adding $Pr(A)$ and $Pr(B)$ from each expert and pool the resulting sums, or we can pool the elicited probabilities for A and B first and then add the pooled results. With a logarithmic opinion pool, these approaches will not yield the same probability, however, it should be noted that this issue does not affect the pooling of survival probabilities.

Another form of expert pooling is the *linear opinion pool*

$$p(\theta) = \sum_{j=1}^{m} w_j p_j(\theta),$$

which is the weighted arithmetic mean of the distributions. This approach is not externally Bayesian. Continuing our example, a weighted sum of Gamma distributions is not a Gamma distribution and is not available in an analytic form unless the rate parameters are equal (Salvo, 2008). Linear pooling does satisfy this marginalisation requirement, however, no aggregation function can simultaneously satisfy the marginalization and externally Bayesian properties.

O'Hagan et al. (2006) notes that when using logarithmic pooling, the decision maker regards as implausible any values of θ that are considered implausible by any single expert. The linear opinion pool, on the other hand, concentrates more in the area where the experts opinions overlap, but it does not rule out values of θ that are supported by only one expert, which may be the reason linear pooling is more commonly used (O'Hagan et al., 2006).

We illustrate these properties in Figure 7.1, in which we consider a hypothetical example where two experts have provided their opinions on θ for an exponential model, with the experts holding somewhat conflicting opinions. We suppose Expert 1 has a prior of $\mathcal{G}(8, 10)$ and Expert 2 has a prior of $\mathcal{G}(20, 10)$. Figure 7.1a presents both pooling approaches for the prior expert opinions, with the purple line representing the density of the opinion obtained by logarithmic pooling while the green refers to the density using linear pooling. As mentioned previously, the logarithmic pooling produces a pooled density that gives most weight to areas of overlap between the expert's opinions, which peaks at the point where the density lines intersect. The linear pool is bimodal and retains the characteristics of the constituent prior distributions, with the 95% credible interval [0.40-2.79] being wider than that of the logarithmic pool (0.77-2.22). Figure 7.1b shows the posterior distributions obtained when the individual expert's priors and the logarithmic and linear pooled distributions are used as a prior for an exponential likelihood with the kernel of a $\mathcal{G}(10, 7)$ distribution. For the logarithmic pooling, when we update each expert's prior with data separately, and then compute the logarithmic opinion pool of these posteriors, or pool both experts' prior opinions and then update with data, we obtain the same posterior distribution (shown by the black line density). In contrast, linear pooling results in two separate posterior distributions depending on whether pooling was conducted on the individual priors (brown/maroon line) or individual posteriors (pink line). It is worth noting that the example presented here relates to the parameter space (i.e. experts gave opinion about the parameter), however, the results also hold when pooling on the opinions in the observable space.

Although expert pooling typically refers to model parameters (i.e. θ) we can also pool expert beliefs about the observable quantity in the same manner. When using our approach as described in Section 7.2.1, when multiple expert opinions are available which we wish to combine with linear pooling, the posterior has the form

$$\pi(\theta|D,\phi) \propto \exp\Big\{ - l\big(g(\boldsymbol{\theta}), \prod_{j=1}^{m} p_j(S(t^*)|\phi_j)^{w_j}, D\big)p(\theta)\Big\},$$

where $\prod_{j=1}^{m} p_j(S(t^*)|\phi_j)^{w_j}$ denotes the pooled distribution for the multiple expert opinions for survival at time point(s) t^* (although for notation ease we have only considered one timepoint). The quantity p_j refers to the (potentially different) probability distributions and associated parameters describing expert $j's$ belief about $S(t^*)$, while for notational ease ϕ_{j} denotes all the information provided by the expert's opinion (i.e. both in terms of parametric models and associated parameters). If we use a linear method to pool our prior information, then the resultant posterior will be different then if we ran separate analyses using each expert opinion and pooled the results *a posteriori*.

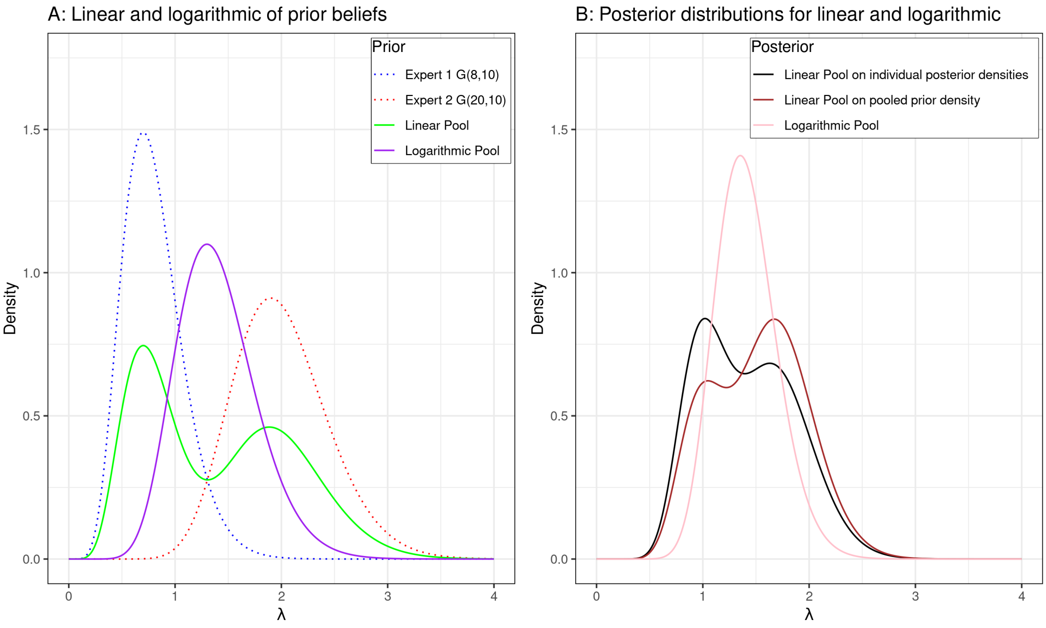


Figure 7.1: Aggregation of expert opinions (with and without data)

7.2.3 Effective Sample Size for Survival Models

One potential approach to determining the ESS of a sample is based on comparing uncertainty in the survival function (elicited from an expert) possibly measured in terms of interval width and match it to a sample size which would produce a similar level of uncertainty. The uncertainty of the Kaplan-Meier survival function is based on a normal approximation to the a binomial distribution.

The standard error of the empirical survival function (assuming no censored observations) is $\text{SE}_{\text{Surv}} = \sqrt{\frac{\hat{S}(t^*)(1-\hat{S}(t^*))}{n_{\text{e}}}}$ and given the probability α and the $z_{\alpha/2}$, the confidence interval $\hat{S}_{\alpha/2}(t^*), \hat{S}_{1-\alpha/2}(t^*)$ is $\hat{S}(t^*) \pm \text{SE}_{\text{Surv}} * z_{\alpha/2}$. Full details on the derivation of the standard error of the survival function are detailed in Collett (2015). If we plug in the median survival probability elicited from the expert as $\hat{S}(t^*)$ and $\hat{S}_{\alpha/2}(t^*)$, $\hat{S}_{1-\alpha/2}(t^*)$ being the lower and upper expert beliefs for a particular probability α we can find an estimate of SE_{Surv} which can in turn be used to find n_e. It should be noted that the estimates of SE_{Surv} and $\hat{S}(t^*)$ could also be the obtained from the standard deviation and median of a parametric distribution fitted to the expert's quantiles. It should be noted that the normal approximation for the above confidence interval can be inappropriate when the interval is close to 0 or 1 and transformations can be considered to produce more accurate intervals (see Collett (2015) for details). Another approach is to recognise that the survival probability can be represented by a $\mathcal{B}(\alpha, \beta)$ distribution. Using the quantiles that were elicited from the expert, a $\mathcal{B}$ distribution can be fit to the expert's opinion and the effective sample size calculated as $\alpha + \beta$.

It is also possible to consider the exponential survival model with a prior on the parameter for which the ESS is directly known. An obvious example is $\lambda \sim \mathcal{G}(\alpha, \beta)$ where α is n_{e}. Using the change of variable technique it is straightforward derive an analytic expression for $S(t^*)$, we can then find the α, β parameters which minimize the difference between the probability distribution assigned to the expert's belief and the distribution implied by α, β through least squares.

Extending this to two parameter survival models, we note that there is no analytical expression for the density of survival at a particular timepoint (as multiple parameters mean that there is no one-to-one transformation). Using a particular prior for the parameters of the Weibull (accelerated time factor) we can introduce a parameter which relates to the effective sample size using the following relations (further details in Bousquet (2006)).

$$
\begin{aligned}
a &\sim \mathcal{U}(0,10)\\
f(a) &= t^{*a}\left(\hat{S(t^*)}^{(-1/n_\mathrm{e})}-1\right)^{-1}\\
\mu|\beta &\sim \mathcal{G}(n_\mathrm{e}, f(a))\\
b &= (1/\mu)^{(1/a)}\\
T &\sim W(b,a)
\end{aligned}
$$

We now have parameters which have an expected survival $\hat{S(t^*)}$ and parameter uncertainty based on n_e. We can simulate a large number of parameters b, a and produce a distribution of $S(t^*)$ from a Weibull W distribution for a range of n_e and find the n_e which is most similar to the expert's distribution (again through least-squares). We also have derived the n_e for the Pareto distribution (see Appendix D.3) and have found that across each of exponential, Weibull and Pareto models that the effective sample size for a given uncertainty level are almost identical.

7.3 Case Study: Inclusion of Elicited Expert Opinion with ELIANA trial

Cope et al. (2019) elicited expert opinion on the expected survival probabilities at 2, 3, 4 and 5 years of pALL patients treated with tisagenlecleucel, based on the available 1.5 year results from the ELIANA trial along with other available tisagenlecleucel data for paediatric acute lymphoma/leukaemia (Grupp et al., 2018, 2015; Maude et al., 2016). Elicitation was conducted in line with the SHELF methodology, in which for each timepoint, experts were asked to first estimate the upper plausible limit (UPL), followed by the lower plausible limit (LPL) so that they are 99% sure that the true survival probability is contained within that interval. Experts were also asked to estimate the most likely values (MLV). A web-based application was developed that would facilitate the elicitation and ensure experts were provided with immediate visual feedback regarding their elicitations, given that information at each timepoint was dependent on that in the previous time point. Before confirming each value, experts were challenged to consider whether they were certain about their estimates, in line with SHELF methodology. Following the individual expert elicitations, consensus about the appropriate long-term survival model from the perspective of a rational impartial observer was achieved in a follow-up meeting (which was the Gompertz model), which allowed experts to interact. A normal distribution was specified using each expert's opinion about expected survival probabilities at each timepoint. The variance of this distribution was determined using the width of the interval provided by the expert. The survival beliefs of the experts implied interval-specific hazards which were incorporated with the ELIANA trial data. Posterior samples for the predicted survival from each expert were pooled to obtain the final survival distribution. In our reanalysis we consider the longer term ELIANA data based on a median duration of follow-up of 24.2 months with range of 4.5-35.1 months Grupp et al. (2018), and combined this with the expert opinions for expected survival for years 4 and 5 (as we have an estimate of the survival function for times $<$ 2.8 years). We considered the expert beliefs at these timepoints and identified which distribution most accurately describes their beliefs, rather than assuming that they were normally distributed. We used the SHELF package to identify the best fitting distribution to these timepoints by minimizing the least square error (Oakley, 2020). Because we wished to include the expert's MLVs, we modified these functions so that the MLV represented the mode of the distribution and included this quantity in the least squares optimization. The best fitting distribution was the one which minimized the sum of squares from either the normal, t, lognormal, gamma or beta distributions. Because we have updated data for survival at years 2 and 2.8 (which we assume is representative of year 3), it is important to confirm that the elicited survival at years 2 and 3 are broadly consistent with the survival at the same timepoints from the updated trial data. For consistency we assumed the same distribution type for each expert across both years, so that the chosen distribution was the one which minimized the total sum of squares across years 2 and 3. The individual distributions for years 2 and 3 are presented in Figure 7.2. Additionally, the logarithm and linear pooling, and a purple interval representing the 95% Kaplan-Meier survival confidence intervals from updated ELIANA data at the same timepoints are plotted.

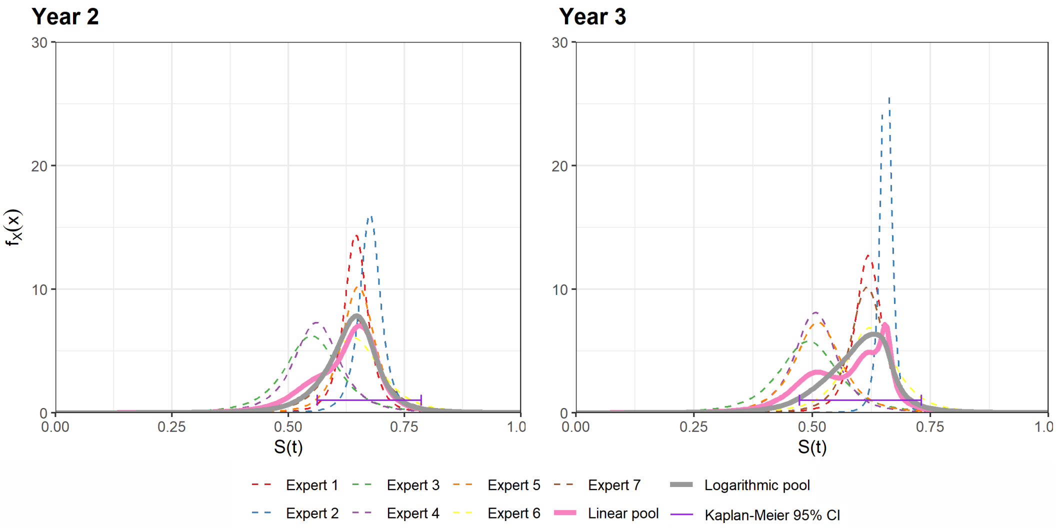


Figure 7.2: Best fitting distributions representing experts' opinion at years 2 and 3 including 95% Kaplan-Meier survival confidence intervals from updated ELIANA data at the same timepoints (purple interval).

Figure 7.2 shows that the pooled distributions have more overlap with the 95% Kaplan-Meier survival confidence interval than the individual expert's distributions and supports the finding that, in general, groups of experts tend to perform better than individuals (Clemen and Winkler, 1999). Although it is probable that experts would re-calibrate their opinions on survival conditional on the longer-term follow-up, based on the observation that the pooled distributions for elicited survival at years 2 and 3 were similar to the follow-up data, it is reasonable to assume that these opinions remain valid, and we incorporate the year 4 and 5 opinions with the updated data.

We repeated the approach described above for the expert opinions at years 4 and 5 with the individual and pooled distributions presented in Figure 7.3. Most of the expert beliefs were described by t distributions with 3 degrees of freedom. Expert 3's opinion is best described using a beta distribution.

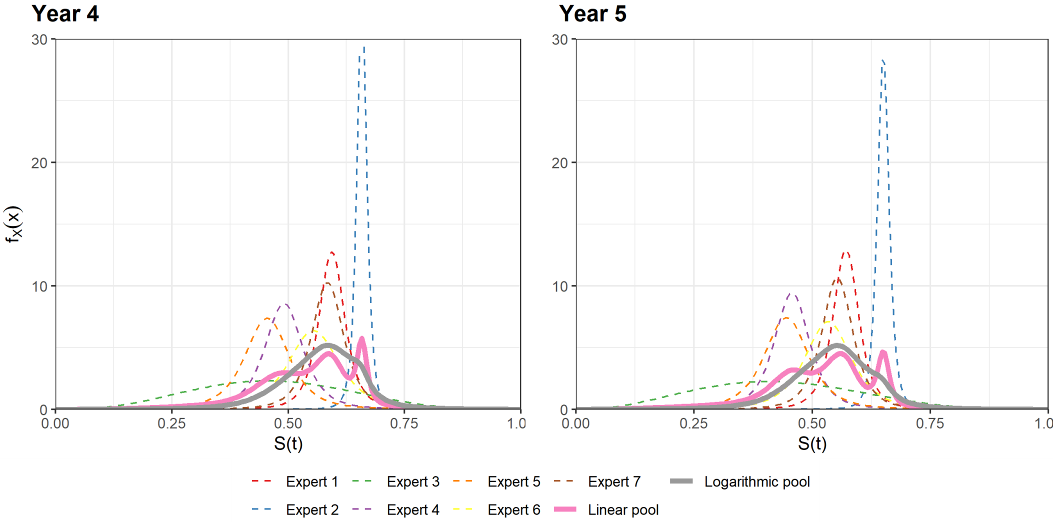


Figure 7.3: Best fitting distributions representing experts' opinion about survival at years 4 and 5.

We see a variety of expert opinions, with Experts 1, 6 and 7 broadly similar, while Expert 4 and 5 are also similar. Across both timepoints Expert 2 has a very strong opinion, while Expert 3 has a diffuse opinion. Although Expert 3 has the widest interval (UPL minus LPL), their MLV is also closer to the LPL, which results in the best fitting distribution having a high standard deviation. Overall this collection of opinions results in a tri-modal distribution for the linear pool. The logarithm pool is smoother and assigns lower probability at the more extreme ends of the parameter space. Because linear pooling is the more common pooling method, we use the linear pooled distributions as representative of the expert opinions which were then incorporated with the updated ELIANA trial data.

7.3.1 Extrapolated Survival

Figure 7.4 shows the predicted survival for the parametric models, including models using the data only (left) and the data together with expert opinion at years 4 and 5 (right). In addition to the posterior median survival for each model, the 2.5% and 97.5% quantiles are presented for the 3 models which have the largest change in 95% interval width at 60 months (Gompertz, Royston-Parmar and generalized gamma). For the models fit with Stan (Stan Development Team, 2020), inference was based on 3 chains each containing 10,000 iterations with the first 5,000 as burn-in, while for models fit with JAGS each chain contained 50,000 iterations and the first 10,000 discarded as burn-in. As shown in Table 7.1 the log-logistic and log-normal models have the best statistical fit with respect to the Deviance information criterion (DIC) (Spiegelhalter et al., 2002)[1]. Models which allowed for rapidly decreasing hazards (Gompertz) or non-monotonic hazards (e.g. log-logistic or log-normal) seem to provide the best fit to the experts opinions and the data, a property which all of the three best fitting models have. However, across all the models considered, the differences in DIC are < 3, suggesting that they are broadly similar in model fit. This is not surprising as the pooled prior is quite diffuse, with a 95% credible interval of [0.28-0.70] for the Year 4 opinion and consequently predicted survival for all parametric models are plausible. If we estimate the models without the expert opinion, the exponential model, which assumes a constant hazard, was the best fit according to DIC. Including expert opinion assigns substantial probability to high long-term survival and the parametric models which accommodate lower long-term hazards fit the data and expert opinion better.

[1] Although our preference is to use WAIC and PML (as in Section 3.7.3), these approaches require us to define the likelihood for each observation, including the contribution for the expert opinion. One approach could be to multiply each observation by $\pi_{t^*}(\theta)^{\frac{1}{n}}$, however, we will use DIC as it does not require us to consider this issue.

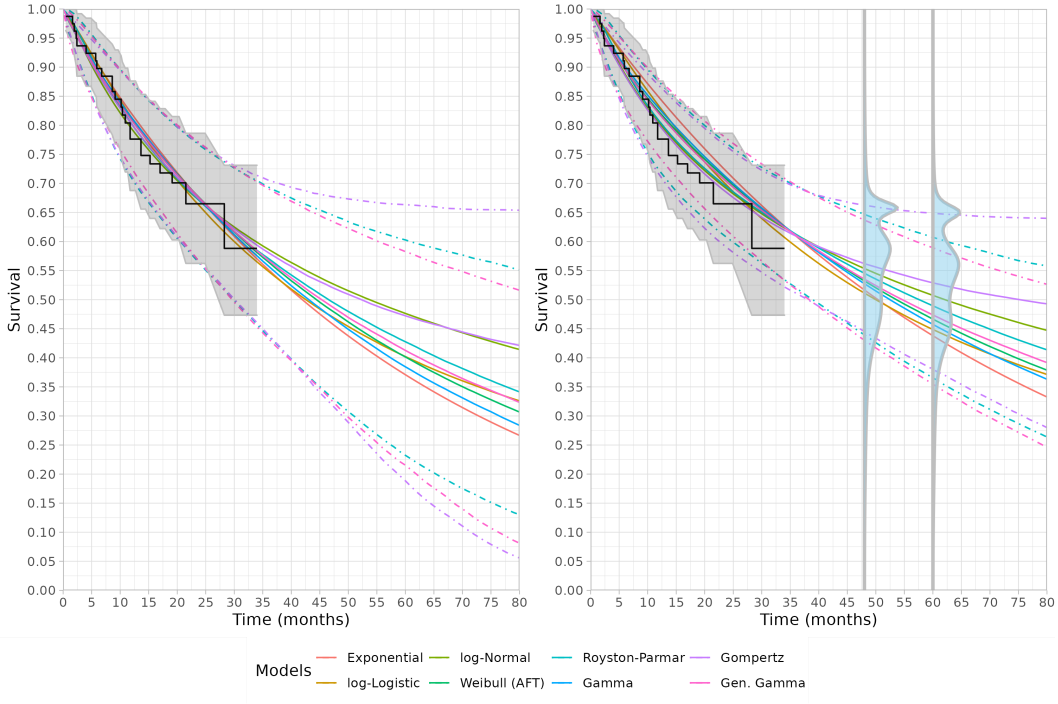


Figure 7.4: Left: Predicted survival functions fit using the updated ELIANA trial without expert opinion. Right: Predicted survival functions using the updated ELIANA trial with expert opinion at 48 and 60 months using linear pooling.

Table 7.1: Survival models ordered by DIC for expert opinion survival models (lower is better)

Models	DIC (expert opinion)	DIC (vague priors)
log-Normal	273.54	277.86
log-Logistic	273.56	277.34
Gompertz	274.05	278.47
Gen. Gamma	274.45	278.55
Exponential	274.63	276.95
Royston-Parmar	275.64	280.64
Weibull (AFT)	275.79	279.19
Gamma	275.83	279.06

7.3.2 Effective Sample Size

We estimated the effective sample size of each of the experts at Year 4 using the four methods described in Section 7.2.3 in Table 7.2. Year 5 results are broadly similar and therefore not presented. We also present the standard deviation of the parametric distribution assigned to the expert's opinion. Because the mode elicited from Expert 2 was not likely to be an accurate representation of the median of the survival, the calculation of the normal approximation to the binomial uses a plug in estimate of SE_{Surv} and $S(\hat{t^*})$ from the best fitting parametric distribution.

Table 7.2: Effective Sample Size of expert opinions including information on most likely value - Year 4 Timepoint

					ESS			
Expert	0.5%	Mode	99.5%	SD	Expo	Wei	Norm	Beta
1	0.43	0.59	0.78	0.05	21	22	57	66
2	0.48	0.66	0.72	0.02	119	120	363	344
3	0.11	0.46	0.86	0.16	5	5	12	9
4	0.19	0.49	0.72	0.07	11	11	27	27
5	0.18	0.45	0.8	0.09	9	9	19	16
6	0.18	0.55	0.87	0.1	5	5	14	11
7	0.28	0.58	0.78	0.06	14	13	37	33

99.5% - Expert upper belief; 0.5% - Expert upper belief; Mode - Mode of expert's belief;
SD - Standard deviation of parametric distribution fit to the expert's belief
ESS - Effective Sample Size under various assumptions;
Expo - Exponential; Weib - Weibull; Norm - Normal approximation;

For Expert 2 the 0.5% quantile is 0.48 and the 99.5% quantile is 0.72, while their MLV is 0.65. This implies quite a heavily skewed distribution which is not well accommodated by any of the distributions we have considered. For the distributions we considered, minimizing the least-squared error of these 3 quantiles only optimized the 99.5% and MLV values with the 0.5% from the fitted distribution being higher than the expert's belief. This resulted in the estimated distributions having too low uncertainty as no distribution can adequately model both the 0.5%, 99.5% with the asymmetric MLV. Because of this we have also presented the results without including the MLV (Table 7.3).

Table 7.3: Effective Sample Size of expert opinions excluding information on most likely value - Year 4 Timepoint

					ESS			
Expert	0.5%	Mode	99.5%	SD	Expo	Wei	Norm	Beta
1	0.43	0.59	0.78	0.07	19	20	53	49
2	0.48	0.66	0.72	0.05	42	42	110	107
3	0.11	0.46	0.86	0.14	4	4	12	9
4	0.19	0.49	0.72	0.1	10	10	23	20
5	0.18	0.45	0.8	0.12	7	7	17	14
6	0.18	0.55	0.87	0.13	5	5	14	11
7	0.28	0.58	0.78	0.1	10	10	26	23

Column Names as per preceding Table.

Overall the ESS based on the parametric models are very similar to each other (i.e. exponential compared to Weibull) and considerably lower than the normal and beta approximations which are also very close in value to each other.

7.4 Including Expert Opinion with Change-point Survival Models

In this section we provide a concise overview on how expert opinion can be incorporated into one of the change-point models presented in Chapter 5. Incorporating expert opinion into these models poses no additional challenges compared to standard parametric models. We examine the combined E1690 & E1684 datasets as outlined in Section 5.5.1.

We include an expert belief that the expected probability of survival of the observation group (OBS) at 15 years is 0.2 with a standard deviation of 0.03 and can be adequately described by a normal distribution. The loss function is the negative log density of the predicted survival of the OBS group at 15 years based on a normal distribution with mean 0.2 and standard deviation 0.03. Fitting a one change-point model (referred to in Chapter 5 as Scenario 1) we obtain the survival function described below in Figure 7.5. Comparing the survival function to the analysis without expert opinion (Figure 5.9), as expected the survival of the OBS arm has fallen. The 95% credible interval for the expected survival probability at 15 years is [0.206 - 0.293], while the survival of the interferon arm (INF) has remained unchanged.

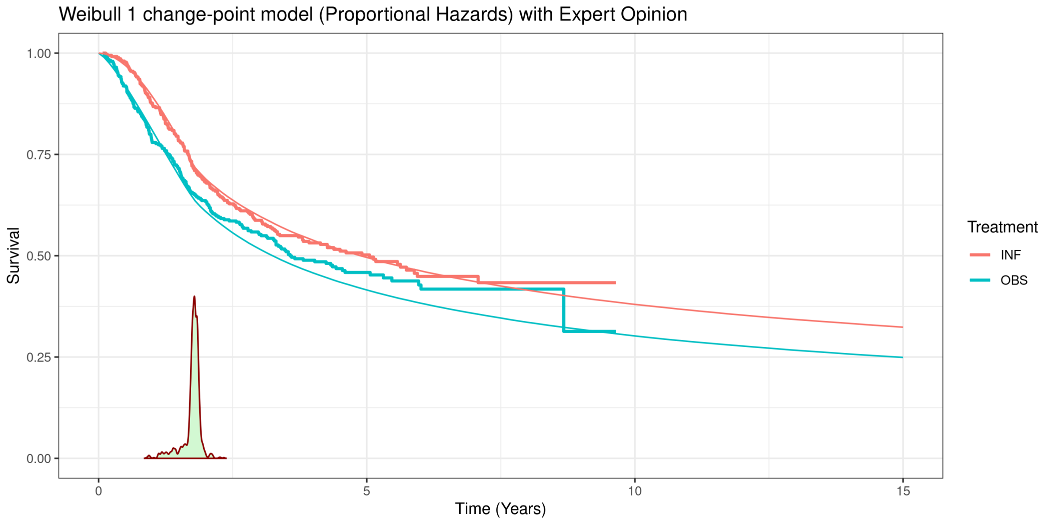


Figure 7.5: Predicted survival function of a one change-point model with expert opinion on survival of the observation arm

7.5 Discussion

The primary contribution of this chapter is that it extends previous work on incorporating expert opinion to a wide range of parametric models (Cope et al., 2019; Ouwens, 2018). In contrast to many previous works, the introduced approach makes it straightforward to incorporate information about other quantities of interest (e.g. median, mean survival, or mean survival difference) into an analysis. The inclusion of expert opinion with a change-point model also links the contributions of Part II of the thesis with Part III, highlighting how it is possible to calibrate the long-term extrapolations of flexible models with expert beliefs.

Additionally, this chapter highlights important considerations with respect to pooling information from multiple experts. Specifically, we describe the estimation of the best fitting probability distributions to each individual opinions, the differences in two pooling approaches and how multi-modal aggregated distributions can be incorporated into the analysis. Our approach permits the use of model selection criteria such as the Deviance Information Criterion (DIC) so that models which have incorporated expert opinion can be objectively compared. Our analysis of the ELIANA trial data results in similar conclusions as the analysis performed by Cope et al. (2019). In their analysis the preferred model was the Gompertz, while in ours the log-logistic ranked highest with both models implying decreasing hazards. In the approach presented by Cope et al. (2019) estimates from each expert were modelled separately, and the overall estimate reflected a combined overall distribution. This necessitated fitting models for each of the individual experts before combining the results. The authors noted that this approach avoids pooling or model averaging, which would provide narrower intervals around the mean. We argue that such an approach does not avoid pooling and is actually a linear pool of the posterior distributions. As our illustrative example in the section on incorporating multiple expert opinions shows, this does not automatically lead to narrower intervals. Decision makers may value an aggregated prior, as it aids understanding about how the prior changes the analysis, compared to an analysis using the data alone. We generally prefer the outcome that we are eliciting to be a single probability distribution representing the combined knowledge of experts in the field (O'Hagan, 2019). Resolving the experts' judgments into a single distribution is known as the problem of aggregation. In this chapter we use mathematical aggregation (as we do not have access to the experts), however, we note that the SHELF framework permits behavioural aggregation in which the group of experts discuss their knowledge and opinions to form "consensus" judgments, to which an aggregate distribution is agreed. Even in situations where behavioural aggregation is the objective, using a mathematical aggregation of the experts' opinions may be a useful visual tool in agreeing the consensus distribution. Although expert opinion can be of value in reducing the differences in extrapolated survival probabilities for different parametric models, the appropriate elicitation of these quantities is challenging. One important point is how the questions are framed, with Bousquet (2006) providing examples of some open questions which are relevant when eliciting beliefs about survival. Clearly defined elicitation questions are particularly relevant as the experts may not be familiar with statistical terms and can misinterpret averages as medians (Bousquet, 2006). It has also been frequently discussed that experts can be overconfident (O'Hagan et al., 2006; O'Hagan, 2019; Lin and Bier, 2008) and that calibration and differential weighting of experts may reduce this overconfidence (Lin and Bier, 2008). Within this analysis it is possible that Expert 2 provided survival estimates that were overconfident, and exclusion of this expert's opinion slightly lowers the expected survival estimates, although the ordering of DIC for the parametric models remains broadly the same, with the Gompertz, log-normal and log-logistic remaining the top three models. When considering the pooled distributions, the 95% intervals of the expert opinions at years 2 and 3 were similar to the 95% intervals from the Kaplan-Meier survival functions at years 2 and 3 for the updated ELIANA data, suggesting that it is appropriate to incorporate the pooled information at years 4 and 5 into our analysis. Because all of the experts had extensive experience in using tisagenlecleucel (or related treatments) in the target population, their pooled opinions can be considered more robust than relying exclusively on the short-term trial data. When the pooled expert opinion was incorporated into the survival analysis, this led to reduced uncertainty in the resultant survival projections. As shown in Figure 7.4, the 95% survival credible intervals for each of the survival models lie within the 95% credible intervals of the expert linear pooled distribution at years 4 and 5. Using model fit statistics will only provide an assessment of fit to the observed data, and a final decision on the choice of model should also be based on clinical plausibility. Incorporating expert opinion in methodologically appropriate ways is therefore a robust way to ensure that decision makers have plausible evidence available to them. Often the plausibility of a parametric model is assessed on the basis that the predicted survival is within an appropriate survival probability interval at a number of landmark timepoints (i.e. between 20-40% at year 5 and 10-20% at year 10). In our opinion, it is best practice to incorporate this information explicitly, and our approach allows for the direct synthesis of these beliefs with the observed data. This approach would be particularly useful in situations where none of the available model projections are considered plausible by decision makers when using data alone, due to e.g., data immaturity, or differences in standards of clinical practice in different countries. If reliable expert opinion is available and can be elicited, our approach permits re-calibration of these models to more accurately reflect the survival projection of the population of interest. Because the opinions elicited from the expert (and parameterized as probability distributions) will almost surely not be centred on the true parameter value (i.e. true survival at a timepoint), it is worth considering for which situations will including expert opinion lead to better estimates of extrapolated survival than using data alone. We explore this in Appendix D.4 through a simulation study and find that in general, if the expert under or overestimates the true survival by $\leq 25\%$ (in relative terms), including expert opinion provides better estimates than using data alone, assuming both the parametric model and data generating process are the same (both Weibull). In the situation where the parametric model chosen was a log-normal and the data generating process was a Weibull distribution, the inclusion of expert opinion produced better extrapolations even when the expert underestimated survival by 40%. We believe that the inclusion of expert opinion can make extrapolation of survival outcomes more reliable, and robust to misspecification of the parametric model. Owing to the number of factors which affect extrapolated survival, further research in this topic is needed.

A number of important points relate to the effective sample size of the expert opinion. We note that the effective sample size from the normal approximation of binomial distribution in estimating uncertainty the survival function and the beta distribution approximation are very similar. This is unsurprising as the beta distribution is related to the binomial distribution. The second key point is that the effective sample size implied by parametric survival models are considerably lower than the normal and beta approximation. This is because parametric survival models, by assuming a functional form to the hazard and survival functions are more "efficient" than the non-parametric estimators. Importantly the different parametric models do not appear to result in different n_e values. It is important to clarify what the value of n_e means in the context of survival models. Given the uncertainty associated with the expert's belief, the n_e suggests that the expert's opinion is equivalent to an equal number of fully observed survival times (i.e. no-censoring before) up until t^*. Additionally as the expert will almost certainly make their assessment with reference to the existing Kaplan-Meier survival function, the n_{e} implicitly takes account of the data and the expert's opinion. The Kaplan-Meier data contains a number of event times n_{events} and it is possible that $n_{\text{events}} > n_{\text{e}}$. In this case the certainty of their opinion about $S(t^*)$ is likely less than the parametric model's uncertainty around $S(t^*)$ using only the data. Even when $n_{\text{events}} > n_{\text{e}}$, the addition of expert opinion to the analysis will almost certainly reduce the uncertainty around $S(t^*)$ as it is effectively further concentrating the $S(t^*)$ estimated by the data only.[2] Given that there are a number of caveats with the interpretation of n_{ESS} it should be interpreted with as a relative rather than absolute measure of informativeness. Estimating n_{e} including or excluding the MLV illustrates that Expert 2 has a very strong belief relative to the other experts ($n_{\text{e}} \approx 42$ assuming a parametric estimator and $n_{\text{e}} \approx 107$ for the non-parametric approach), especially as even in the updated data $n_{\text{e}} = 25$ (with number of events at original data-cut off not available). It is possible that experts are not familiar with 99% belief intervals as 95% intervals are much more common in statistical literature. Therefore, this expert could have underestimated that the 99% intervals should be ≈ 1.31 times wider than their 95% interval.

Although it is not apparent which estimate of n_{e} should be presented to the expert, both illustrate evidence of overconfidence in this expert's survival estimates. We believe that because informing the extrapolation of the survival is the objective of the elicitation, we should present n_{e} from the parametric results. It is important to highlight that the n_{e} at multiple timepoints are not additive. Survival probabilities at the multiple timepoints are not independent and it may be most appropriate to present the n_{e} for only one timepoint. More generally, including expert opinion at multiple timepoints by way of the loss function assumes that each of these pieces of information is independent. This assumption is not appropriate especially if there are multiple timepoints very close together and will result in underestimation of uncertainty. One potential solution might be to elicit an opinion on the survival at a particular timepoint and then for all future timepoints the probability of surviving to that timepoint conditional upon surviving to the previous timepoint.

Although not discussed in this chapter, there are situations where the expert may have considerable experience with the comparator arm and may be more comfortable providing an opinion on the plausible survival probabilities for the comparator at particular timepoint(s). If a relationship such as proportional hazards (PH) or accelerated time factor (ATF) can be considered tenable (i.e. evaluated based on trial data and assumed to hold in the long-term), a survival model with the PH or ATF parameterization with treatment status as a covariate could be estimated. Alternatively, experts may be willing

[2] There are some trivial potential situations which expert opinion might increase uncertainty in the posterior survival.

to provide an estimate of the expected survival difference between two treatments. Both of these approaches have been implemented and we provide simulated examples of each situation described in Part IV. As noted, similar results can be obtained using frequentist methods (although this would not be the case for the multi-modal expert opinions) and `expertsurv` provides code based on the `flexsurv` package (Jackson, 2016) to accommodate this. Although the incorporation of expert opinion is relatively straightforward with the approach described in this chapter, further research on elicitation of long-term survival probabilities and best practices are important if expert opinions are to become more widely used in health technology assessments using time-to-event outcomes.

Part IV

Software Applications

8 Contributions for R in HTA

8.1 Overview of PiecewiseChangepoint package

The goal of the PiecewiseChangepoint package is to provide a suite of R functions to estimate the number and locations of change-points in piecewise exponential models described in Chapter 3. In order to efficiently estimate these models we enhanced the computational performance by rewriting key functions in C++ using the Rcpp package (Eddelbuettel and François, 2011). During the model estimation we do a Gibbs step for each change-point location which requires us to evaluate the marginal likelihood at each event time in a given interval and which is computationally intensive. Using C++ allows us to reduce the following computational bottlenecks:

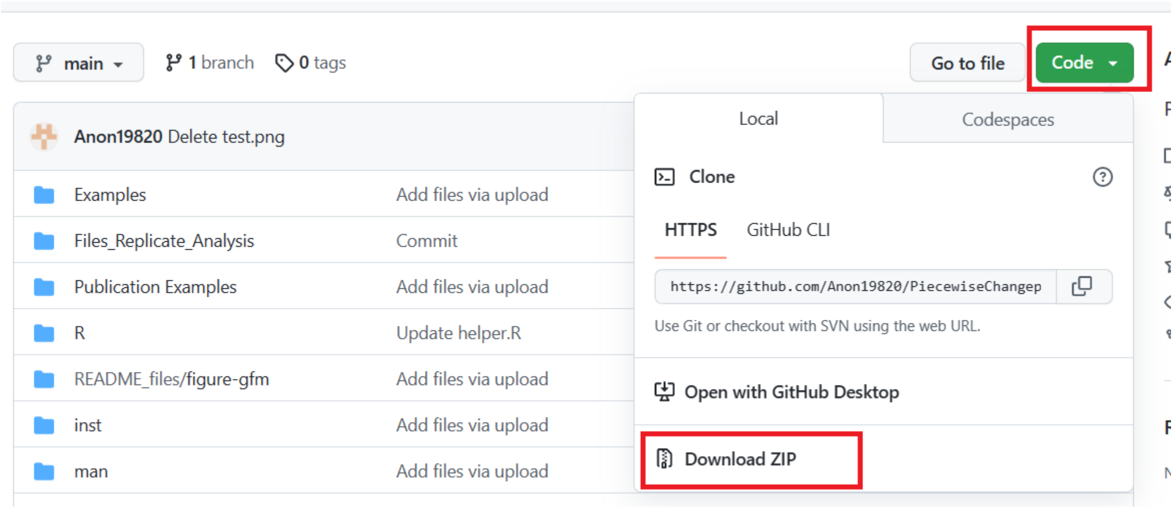
- Loops that can't be easily vectorised because subsequent iterations depend on previous ones.
- Problems which involve calling functions many times as the overhead of calling a function in C++ is much lower than that in R.

Both of the above points are by default present within the implementation of our code i.e., due to the recursive evaluation of the marginal likelihood and for the number of simulations required to ensure the sampler has fully explored the posterior distribution. As a result re-writing key sections of the code in C++ resulted in a 10 times optimization of the model running time.

A full overview of Rcpp is provided by Eddelbuettel (2013) with details on how to implement Rcpp code with a R package is provided in Chapter 25 of Advanced R (Wickham, 2019).

8.1.1 Installation

You can install the binary version of the package by downloading the PiecewiseChangepoint_1.0.zip file within the main Github folder (https://github.com/Anon19820/PiecewiseChangepoint). This file and any other which are hosted on Github can be individually downloaded, however, it may be more convenient to download all files by accessing the main Github folder, select "Code" and then "Download Zip" (Figure 8.1). Once this folder is unzipped all files from the repository are available.



Anon19820 Delete test.png

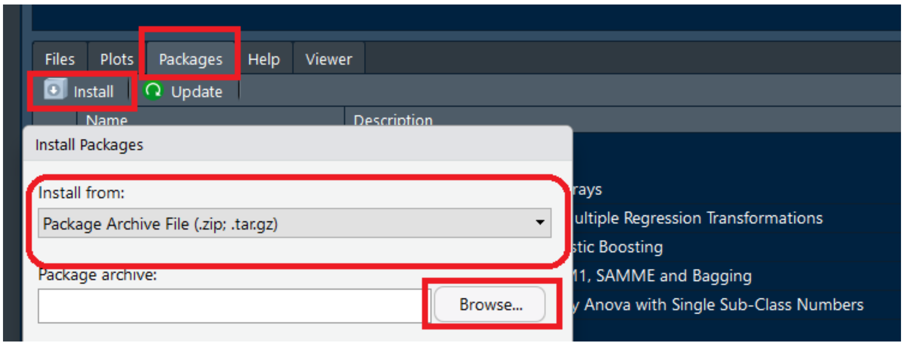
Examples	Add files via upload
Files_Replicate_Analysis	Commit
Publication Examples	Add files via upload
R	Update helper.R
README_files/figure-gfm	Add files via upload
inst	Add files via upload
man	Add files via upload

https://github.com/Anon19820/PiecewiseChangep

Use Git or checkout with SVN using the web URL.

Figure 8.1: Downloading all package files from Github

Within R open the Packages window, select the "Install" button within the "Install from:" drop-down and select "Package Archive File (.zip; .tar.gz)" and find the PiecewiseChangepoint.zip file with "Browse" button (Figure 8.2).



Install Packages

Install from:
Package Archive File (.zip; .tar.gz)

Package archive:

rays
ultiple Regression Transformations
stic Boosting
11, SAMME and Bagging
y Anova with Single Sub-Class Numbers

Figure 8.2: Installation of binary package from Github

Alternatively if you have the remotes package by Csárdi et al. (2023) installed you can install the package directly from Github with:

```
remotes::install_github("Anon19820/PiecewiseChangepoint")
```

After the successful installation, running `library`(‘‘`PiecewiseChangepoint`’’) will provide access to the functions described in the subsequent sections. In order to run the function `compare.surv.mods`, we require JAGS and Stan software programs (Plummer, 2003; Stan Development Team, 2020). Additionally we require the R packages `rjags`, and `R2jags` (Plummer, 2022; Yu-Sung and Masanao, 2015) to help with the processing of the MCMC samples.

8.1.2 Simulated Example

First we load the package and simulate some piecewise exponential data.

```
## simulated example
set.seed(123)
n_obs =300
n_events_req=300
max_time =  24 #months
rate = c(0.75,0.25)/12 #we want to report on months
t_change =12 #change-point at 12 months
df <- gen_piece_df(n_obs = n_obs,n_events_req = n_events_req,
                   num.breaks = length(t_change),rate = rate ,
                   t_change = t_change, max_time = max_time)
```

We see the output of this dataframe below:

```
head(df)
    time_event status       time
    0.09194727      1 0.09194727
    0.23141129      1 0.23141129
    0.24251702      1 0.24251702
    0.25450622      1 0.25450622
    0.28833655      1 0.28833655
    0.32615105      1 0.32615105
```

For this simulated dataset; `time_event` represents the time the event would occur at in the absence of censoring, while `time` is the minimum of the censoring time and the event time. The column named `status` is an indicator variable if the event occurred at the corresponding time or if it was censored. Plotting the survival function (Figure 8.3) we see a potential change in the hazard at around Year 1.

```
# Fitting survival models - requires survival package
require("survival")
# Drawing survival functions - requires surviner package
require("survminer")

ggsurvplot(fit, palette = "#2E9FDF")
```

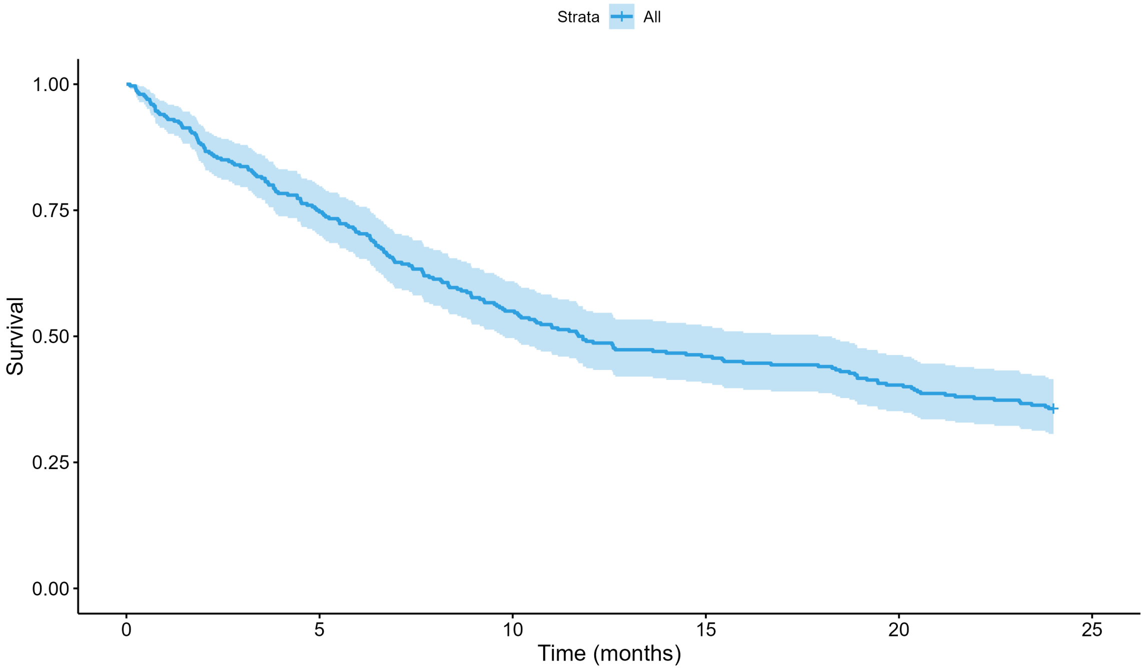


Figure 8.3: Empirical survival function for simulated data

As noted by Bagust and Beale (2014), constant hazards are linear with respect to the cumulative hazard function, therefore, the change in hazards at approximately 12 months can be seen more clearly in Figure 8.4.

```
ggsurvplot(fit, palette = "#2E9FDF", fun = "cumhaz")
```

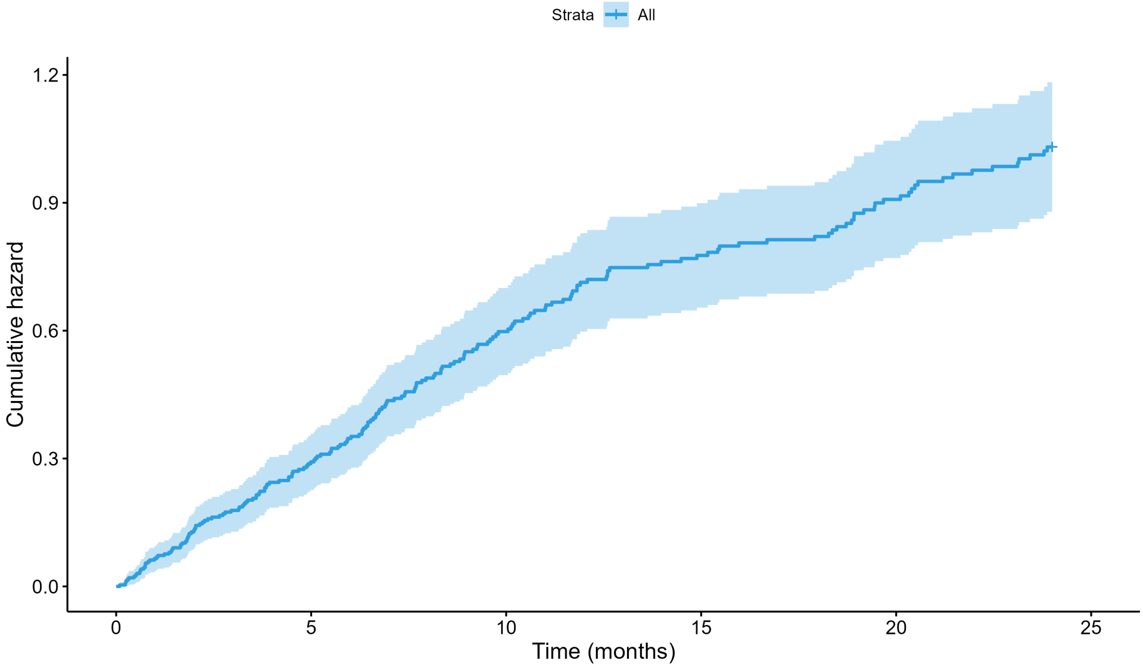


Figure 8.4: Empirical cumulative hazard function for simulated data

Next, we fit the piecewise exponential model noting that only the time and status columns are required. The timescale argument changes the prior for the hazards λ so that it is appropriate for the timescale. For example, if the timescale is years then the a vague prior centered around 1 is appropriate (i.e. 36% of population having the event each year), while if the timescale is in months the equivalent prior should have an expected value of 1/12 (and days 1/365).

```
#timescale can be also equal to "years" or "days"
Collapsing_Model <- collapsing.model(df,
                                     n.iter = 20750,
                                     burn_in = 750,
                                     n.chains = 2,
                                     timescale = "months")
```

As we would expect the one change-point model has the highest posterior probability.

```
Posterior Change-point Probabilities:
       0         1         2         3         4         5
0.000375  0.808075  0.166975  0.021900  0.002250  0.000425

Summary of 1 change-point model:

  changepoint_1      lambda_1           lambda_2
  Min.   : 0.8179    Min.   :0.04148    Min.   :0.01301
  1st Qu.:11.7006    1st Qu.:0.05620    1st Qu.:0.02356
  Median :11.9004    Median :0.05941    Median :0.02652
  Mean   :11.8586    Mean   :0.05954    Mean   :0.02682
  3rd Qu.:12.6027    3rd Qu.:0.06271    3rd Qu.:0.02978
  Max.   :15.9675    Max.   :0.09392    Max.   :0.05050
```

Simulations from the posterior distribution for the change-point locations and associated hazards can be extracted from the returned object using the "$" (highlighted in the code below).

```
Collapsing_Model$changepoint
Collapsing_Model$lambda
```

Once we run the model long enough (20,000 simulations over 2 chains should be more than enough), we may want to look at a plot of the survivor function. In health economics we are typically interested in long-term survival of parametric models. In this situation we want a plot of the first 60 months which we can do using the `max_predict` argument (in this case 60 months). The red lines show the individual posterior simulations and are a natural representation of the parameter uncertainty. The grey lines show the 95% credible interval for the survival function (Figure 8.5).

```
plot(Collapsing_Model, max_predict = 60)+xlab("Time (Months)")
```

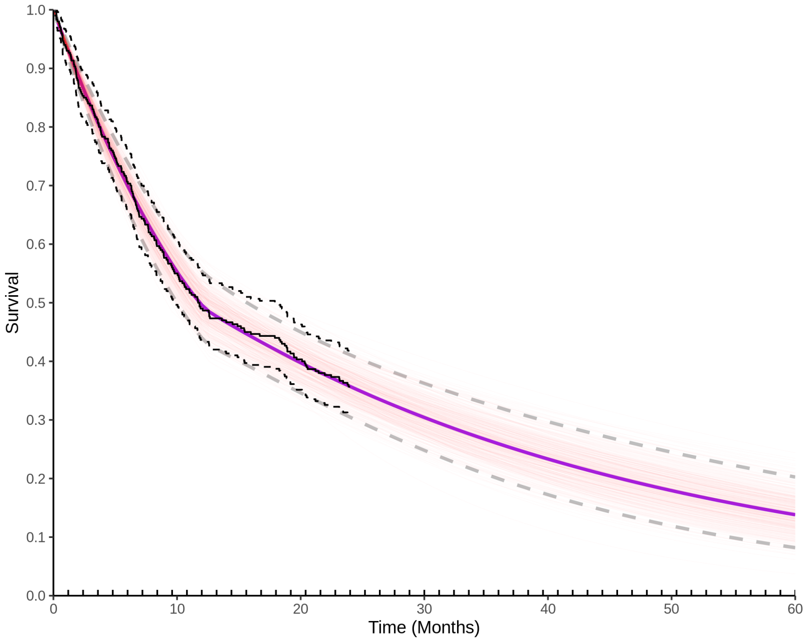


Figure 8.5: Predicted survival function from Piecewise Exponential Model

Similarly, we may also want to look at the hazard function. In this situation we only present the hazard up to the maximum time observed in the data. This is because by definition, the hazard from the final interval will be the one which is extrapolated throughout the time horizon (in a later section we will adjust this hazard for general population mortality).

```
plot(Collapsing_Model, type = "hazard")+xlab("Time
↪  (Months)")+ylab("Hazards")+ylim(c(0,.1))
```

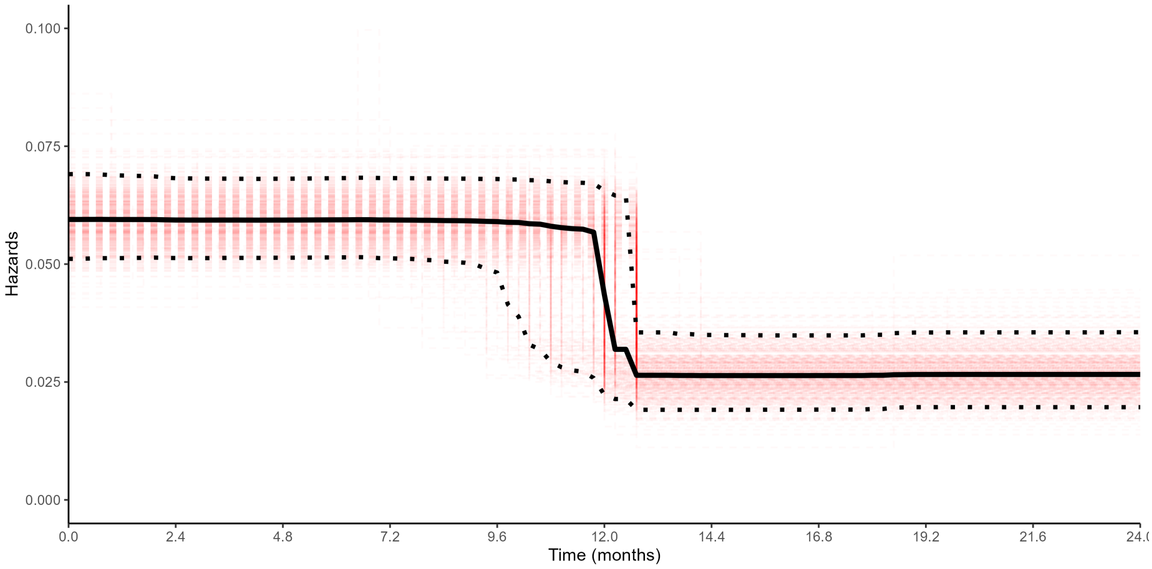


Figure 8.6: Predicted hazard function from Piecewise Exponential Model

By default, the plot methods described above use all the posterior simulations. If for example, we were only interested only in the 1 change-point model, we can specify this using the chng.num argument. The green diamonds indicate the mean location of the change-points. When plotting chng.num = "all" (default) of the simulations there is no sensible mean location of the change-points as there are different numbers of change-points and they are therefore not plotted.

```
plot(Collapsing_Model, max_predict = 60, chng.num = 1)+xlab("Time (Months)")
```

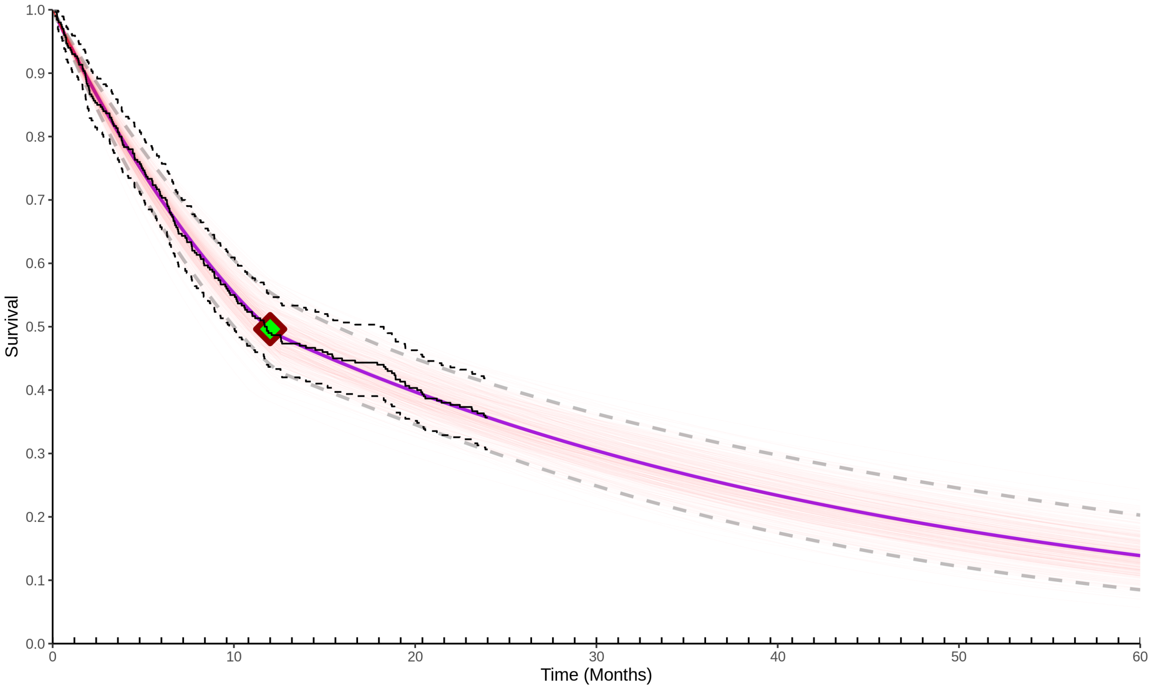


Figure 8.7: Predicted survival function from one change-point piecewise exponential model

In practical health economic modelling, we require evaluation of the survival function. Using the notation from Chapter 3 and assuming time t (at which we calculate the survival probability) is within the j^{th} interval, we calculate the cumulative hazard for this interval as $\lambda_j(t - \tau_{j-1})$ with τ_{j-1} being the $j-1^{\text{th}}$ change-point and λ_j the j^{th} hazard. We also require the cumulative hazard for all the previous intervals which is $\sum_{g=1}^{j-1} \lambda_g(\tau_g - \tau_{g-1})$. As can be seen from Equations 2.6 and 2.7 the survival probability is the exponential of the negative of the total cumulative hazard and is written fully as:

$$S(t) = \exp\left\{ - \left[\lambda_j(t - \tau_{j-1}) + \sum_{g=1}^{j-1} \lambda_g(\tau_g - \tau_{g-1})\right]\right\}.$$

Within the R package the function `get_Surv` evaluates the survival at a vector of user specified times. The user can specific an additional argument `chng.num` if they require survival probabilities from a particular change-point number.

```
St_all <- get_Surv(Collapsing_Model, time = c(0:60))
St_all <- get_Surv(Collapsing_Model, time = c(0:60), chng.num = 1)
```

Because economic models are primarily developed in Microsoft Excel, we have written a Visual Basic Application (VBA) function to calculate the survival probability for the PEM (Code Chunk 7). This function is implemented with an example in an excel file called VBA_PEM.xlsm.

```
Function Survival_PEM(time, params, changepoints, n_interval As Integer) As
↪   Double
Application.Volatile
   Cum_Haz = 0
   Dim changepoints2() As Double
   ReDim changepoints2(1 To n_interval + 1) As Double
   changepoints2(1) = 0

      For i = 2 To n_interval
        changepoints2(i) = changepoints(i - 1)
      Next i

   changepoints2(n_interval + 1) = 99999

   For i = 2 To (n_interval + 1)
     If time > changepoints2(i - 1) And time <= changepoints2(i) Then
        Cum_Haz = Cum_Haz + (time - changepoints2(i - 1)) * params(i - 1)
     ElseIf time > changepoints2(i) Then
        Cum_Haz = Cum_Haz + (changepoints2(i) - changepoints2(i - 1)) *
        ↪   params(i - 1)
     Else
       Cum_Haz = Cum_Haz
     End If
   Next i

Survival_PEM = Exp(-Cum_Haz)
End Function
```

Listing 7: PEM Visual Basic Application Code

8.1.3 Including General Population Mortality

Including General Population Mortality (GPM) is required to ensure that the extrapolated hazards are consistent with the increasing hazards associated with advanced ageing. Adjustments for GPM are typically done within the cost-effectiveness model; however, we can include them directly at the analysis stage so that we see their impact on the extrapolated survival. In this section we consider different approaches to including general population mortality with the piecewise exponential change-point model estimated using the simulated

Data Sources and approaches to including GPM

We consider GPM from a UK data source which provides mortality rates, defined as “the probability of that a person aged exactly x will die before reaching $x + 1$” (Office for National Statistics , ONS). Therefore, this data source provides the conditional probability of death within a year at each age.

Assuming our population is 50% male and female and the age at baseline is 55 years we have the following conditional probabilities of death at each age:

```
age_baseline_example <- 55
prop_male <- 0.5
time_horizon <- 100
Conditional_Death_df <- read.xlsx("Conditional_Death_UK.xlsx", 1) %>%
↪  filter(age >=age_baseline_example)
colnames(Conditional_Death_df) <- c("age", "Male_cond_death",
↪  "Female_cond_death")
head(Conditional_Death_df)
```

There are a number of approaches which can be used to incorporate these gender specific mortality probabilities with the hazards generated from the parametric model and is discussed in detail elsewhere (van Oostrum et al., 2021). Typically a cohort approach is considered, whereby for each timepoint after baseline the age of the cohort is simply the average age at baseline plus the time since the baseline. Furthermore, the proportion of males vs females is assumed to remain constant. This constant proportion assumption is not technically correct as males and females are subjected to different general population mortality so that at older ages we expect a greater proportion of females. To calculate this dynamic proportion, for each age after the average age at baseline we calculate the probability of survival to that age for both male and female (conditional on survival at the baseline age). These probabilities are used to calculate the proportion of males/females surviving at each timepoint. We refer to these approaches as “cohort static/dynamic gender proportion approach".

Both the cohort approaches described above assume that the general population survival estimated at the population averages of age and gender are similar to the average of the general population survival estimated across the trial population, something which is not guaranteed to be the case (Briggs et al., 2006). In this situation, we might consider incorporating each individuals’ specific age/gender general population mortality with the posterior distribution for survival predicted from the parametric model to obtain predicted survival which is then averaged, which we term “simulation approach”.

One other approach we shall discuss is the incorporation of the general population mortality using the “internal additive approach”. Rather than including general population mortality in a post-hoc manner it is possible to include it directly into the likelihood of the model. The hazard function now decomposes all-cause mortality (ACM) into disease-specific/excess mortality (DSM) and GPM by adding GPM hazards in the log-likelihood function of a parametric distribution.

Modelling the hazard function using the internal additive approach may yield different results to other approaches if there is an appreciable level of general population mortality within the trial period. This is because the other approaches to incorporating GPM assume that the parametric model fit to the data only estimates the disease-specific mortality (with the implicit assumption that GPM is minimal within the trial period). Secondly, the elevated within trial GPM implies that the age distribution in the risk set at later timepoints in the trial would be different than that which is implied by the cohort assumption (i.e., average age at baseline plus time since baseline) used in the other methods. It is unclear how frequently these factors will result in an appreciable difference in predicted survival and this approach does not appear to be commonly employed (van Oostrum et al., 2021). In the examples where we extrapolated survival (and therefore required to include general population mortality) the average ages at baseline were between 55-65, therefore, the general population mortality during the follow-up period in our trials would be low, suggesting the internally additive hazard approach is unlikely to produce results which are different to the other approaches for incorporating general population mortality. Furthermore, because we do not have patient level data for the examples presented in the manuscript, we cannot directly test this assumption.

Derivation of general population hazards

Irrespective of the approach taken we are required to convert the conditional death probabilities to rates. In our example the timescale is months and we need to convert this annual probability to a monthly rate which is done using the following formula (Fleurence and Hollenbeak, 2007) (assuming a constant rate of mortality):

$$r = -\frac{1}{t}\ln(1-p).$$

Because there are 12 months in a year, $t = 12$ and p is the specific (in our case annual) probability of death.

With the below R code we now have the monthly rate of death for ages 55 (our assumed starting age of the cohort) up to 100 years of age, adjusted for distribution of males and females (which is can either be static or dynamic).

```
time_factor <- 12
df_temp <- Conditional_Death_df
#Code to get the get convert probability to hazards
df_temp$Male_cond_haz <- -log(1-df_temp$Male_cond_death)/time_factor
df_temp$Female_cond_haz <- -log(1-df_temp$Female_cond_death)/time_factor
df_temp <- df_temp %>% filter(age >= age_baseline_example & age <=
↪   time_horizon)
n_row_df_temp <- nrow(df_temp)
time_partial_vec <- rep((0:(time_factor-1))/time_factor)
df_temp <- do.call("rbind", replicate(time_factor, df_temp, simplify =
↪   FALSE)) %>%
  arrange(age)
df_temp$age <- df_temp$age+  rep(time_partial_vec,times = n_row_df_temp)

#Cohort - Static Gender proportion
df_temp[, "mix_haz_static"] <- df_temp[,"Male_cond_haz"]*prop_male +
↪   df_temp[,"Female_cond_haz"]*(1-prop_male)

#Cohort - Dynamic Gender proportion
df_temp$male_St <- exp(-cumsum(df_temp$Male_cond_haz))
df_temp$female_St <- exp(-cumsum(df_temp$Female_cond_haz))

df_temp$prop_male <- df_temp$male_St/(df_temp$male_St +df_temp$female_St)
df_temp[, "mix_haz_dynamic"] <- df_temp[,"Male_cond_haz"]*df_temp$prop_male
↪   + df_temp[,"Female_cond_haz"]*(1-df_temp$prop_male)
#Assume the time at baseline is time zero and subject to no hazard
gmp_haz_vec_example = df_temp[-1, "mix_haz_static"]

#Creates a data.frame of GPM hazards which is used in compare.surv.mods
↪   function
gmp_haz_df_example <- data.frame(time = 1:length(gmp_haz_vec_example),
hazard = gmp_haz_vec_example)
```

Within the `compare.surv.mods` function (described in more detail in Section 8.1.4) the cumulative hazard of death (associated with GPM) and cumulative hazard of an event (from the parametric model) is added to obtain the overall cumulative hazard $H(t)$. The cumulative hazard is the sum (in the case of discrete hazards as above) of the individual hazards and the integral of the parametric hazards. The dataset called `gmp_haz_df_example` containing the GPM hazards generated in the above code is supplied to the `gmp_haz_df` argument in the `compare.surv.mods`. By default the `compare.surv.mods` function only implements GPM hazards after the extrapolated portion of the data (as we observe survival from all causes up until then), although GPM hazards can be added from start of follow-up by using the `gpm_post_data = FALSE`. It should be noted that only cohort static or dynamic gender proportion approaches can be used with the `compare.surv.mods` function.

We see in Figure 8.8 that including the GPM hazard ensures that the extrapolated hazard exhibits the characteristic increasing hazards associated with ageing.

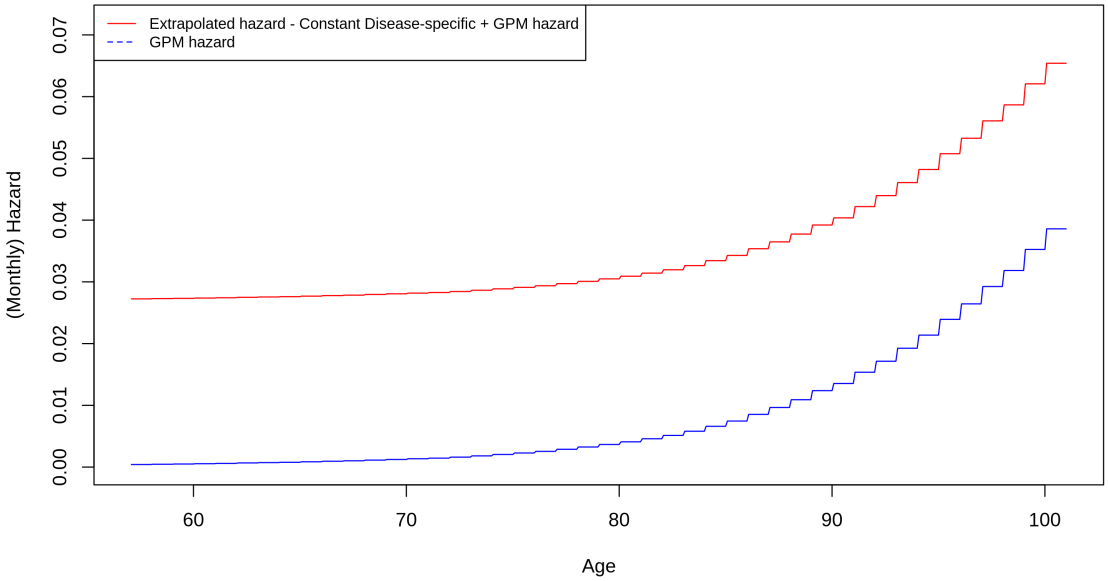


Figure 8.8: Predicted hazard function by age including a constant disease specific hazard

Implementation of the other two approaches (simulation and internal additive approach) are considered in a separate R script titled `GPM_examples_final.R` within the `Files_Replicate_Analysis` folder. For the "internal additive approach" we estimate the piecewise exponential model using a custom written JAGS script rather than the model described in Chapter 3 as the model likelihood now includes the non-standard contribution of the GPM. The model written in JAGS is substantially more computationally intensive and requires the number of change-points (but not their locations) to be fixed. This is a minor limitation as we can use the approach described in Chapter 5 to find the most probable change-point model and then fit that change-point model with the custom JAGS script (or alternatively fit several change-point models and consider the one with the lowest WAIC or other goodness of fit statistic).

The key difference in this model is that both the log hazard function and cumulative hazard function which enter the log-likelihood of the model require the addition of a term for the the general population hazard $h_{\text{GPM}}(t_i)$ and the general population cumulative hazard $\Lambda_{\text{GPM}}(t_i)$ at each individual's event or censoring time. Equation 3.2 is adjusted so that the individual contribution of the log hazard function is $\sum_{j=1}^{k+1} \delta_{ij} v_i (\log(\lambda_j + h_{\text{GPM}}(t_i))$ and the cumulative hazard function is $\sum_{j=1}^{k+1} \delta_{ij} \left[\lambda_j (t_i - \tau_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (\tau_g - \tau_{g-1}) + \Lambda_{\text{GPM}}(t_i) \right]$. The general population hazard and cumulative hazard values are fixed and not model parameters, determined by the GPM data source, gender and age and the time of event/censor.

We implemented four approaches (cohort static and dynamic gender proportion, simulation and internal additive approaches) to including general population mortality with the same parameters as used in the simulated dataset from Section 8.1.2. The proportion of males was 50% and the mean age at baseline was 55 with a standard deviation of 10 and a maximum age of 90. Survival times were adjusted to account for the impact of increased age on survival time. As shown in Figure 8.9 all the approaches produced very similar extrapolations. Increasing the average age at baseline to 75 still resulted in very similar extrapolations across the approaches (not shown).

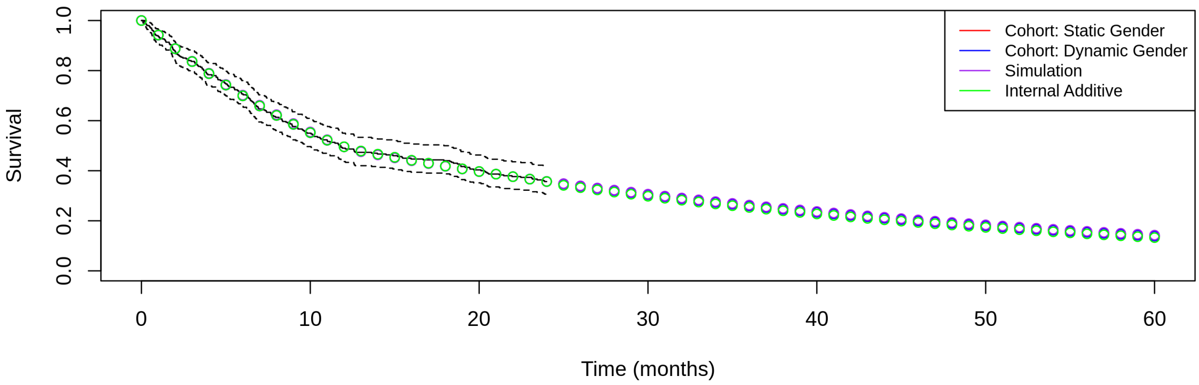


Figure 8.9: Extrapolated survival using different methods to incorporate general population mortality

8.1.4 Fitting of Standard Parametric models and Plot of Extrapolated Survival

In health economics we are typically interested in picking between one of a number of alternative parametric survival models, although it is also possible to combine survival functions from all models using model averaging (Jackson et al., 2010). We can compare the piecewise exponential model with seve commonly used parametric models along with Royston-Parmar spline models. We fit the models using JAGS and Stan (Plummer, 2003; Stan Development Team, 2020) and compare the model fit using Widely Applicable Information Criterion (WAIC) (Watanabe, 2010).

The fitting of other parametric models is accomplished by the `compare.surv.mods` and general population mortality is adjusted for by including a `gmp_haz_df_example` as described above. Fitted models include exponential, Weibull, gamma, Gompertz, log-normal, log-logistic, generalized gamma and Royston-Parmar cubic splines (choice between one and two knot models based on lower WAIC). Model fit to the observed data and a plot of the extrapolated survival are available from within the `mod_comp` object along with the posterior samples from all the fitted models.

```
#Below function can take a number of minutes to evaluate
set.seed(123)
mod_comp <- compare.surv.mods(Collapsing_Model,
                              max_predict = 100, #100 months
                              n.iter.jags = 5000, #Run JAGS/Stan for 5000
                              ↪   samples
                              n.thin.jags = 1,
                              n.burnin.jags = 500,
                              chng.num = 1, #Using results from 1
                              ↪   change-point PEM
                              gmp_haz_df =gmp_haz_df_example) #GPM dataset

#Returns a dataframe with the model fit results
mod_comp$mod.comp[,c(1,3)] %>% arrange(WAIC)
```

```
                  Model    WAIC
1 Piecewise Exponential 1547.59
2            Log-Normal 1552.32
3          Log-Logistic 1553.21
4              Gompertz 1553.25
5 Royston-Parmar 2 knot 1553.76
6      Generalized Gamma 1556.38
7               Weibull 1561.83
8                 Gamma 1564.01
9           Exponential 1568.01
```

```
#Returns a Survival plot with PEM and 3 best fitting models
mod_comp$plot_Surv_all
```

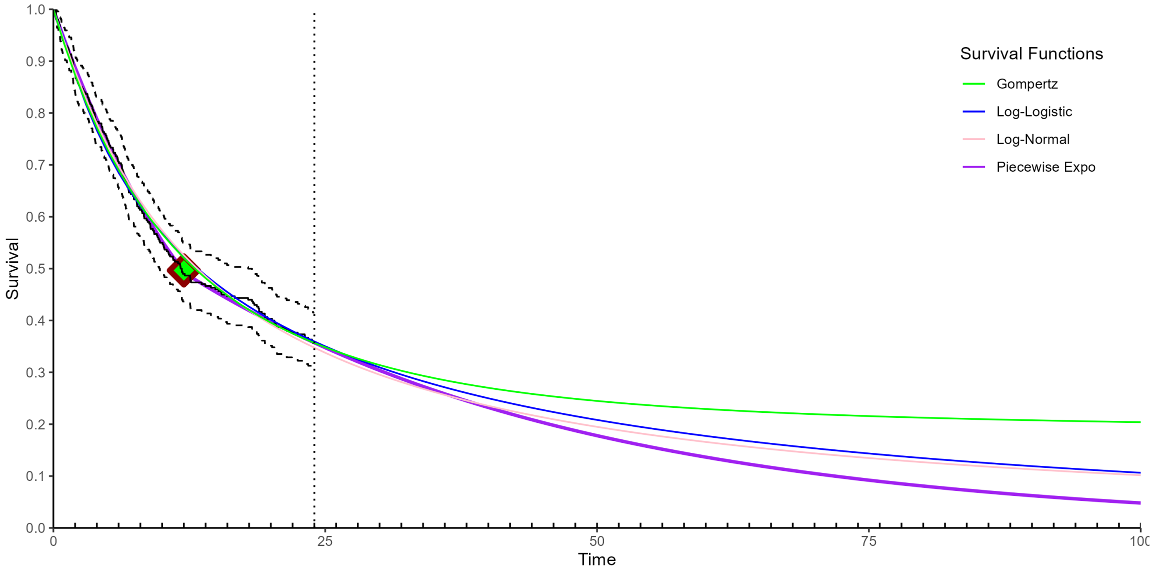


Figure 8.10: Predicted survival for piecewise exponential model and three best fitting other parametric models

8.2 expertsurv R-package

The goal of the R package `expertsurv` is to incorporate expert opinion into an analysis of time-to-event data. The package uses many of the core functions of the `survHE` package (Baio, 2020).

The key function is `fit.models.expert` and operates almost identically to the `fit.models` function of `survHE`.

8.2.1 Installation

You can install the latest version of expertsurv from GitHub with:

```
devtools::install_github("Anon19820/expertsurv")
```

or from CRAN directly:

```
install.packages("expertsurv")
```

8.2.2 Expert Opinion on Survival at timepoints

If we have elicited expert opinion of the survival probability at certain timepoint(s) and assigned distributions to these beliefs, we encode that information as follows:

```r
#A param_expert object; which is a list of
#length equal to the number of timepoints
param_expert_example1 <- list()

#If we have 1 timepoint and 2 experts
#dist is the names of the distributions
#wi is the weight assigned to each expert (usually 1)
#param1, param2, param3 are the parameters of the distribution
#e.g. for norm, param1 = mean, param2 = sd
#param3 is only used for the t-distribution and is the degress of freedom.
#We allow the following distributions:
#c("normal","t","gamma","lognormal","beta")

param_expert_example1[[1]] <- data.frame(dist = c("norm","t"),
                                         wi = c(0.5,0.5), # Ensure Weights
                                         ↪   sum to 1
                                         param1 = c(0.1,0.12),
                                         param2 = c(0.005,0.005),
                                         param3 = c(NA,3))
param_expert_example1
#> [[1]]
#>   dist  wi param1 param2 param3
#> 1 norm 0.5   0.10  0.005     NA
#> 2    t 0.5   0.12  0.005      3

#Naturally we will specify the timepoint for which these probabilities where
↪   elicited
timepoint_expert <- 14
#In case we wanted a second timepoint -- Just for illustration

# param_expert_example1[[2]] <- data.frame(dist = c("norm","norm"),
#                                          wi = c(0.5,0.5),
#                                          param1 = c(0.05,0.045),
#                                          param2 = c(0.005,0.005),
#                                          param3 = c(NA,NA))
#
# timepoint_expert <- c(timepoint_expert,18)
```

If we wanted opinions at multiple timepoints we just include append another list (i.e. `param_expert_example1[[2]]` with the relevant parameters) and specify `timepoint_expert` as a vector of length 2 with the second element being the second timepoint.

For details on assigning distributions to elicited probabilities and quantiles see the `SHELF` package (Oakley, 2020), and for an overview on methodological approaches to eliciting expert opinion see O'Hagan (2019). We can see both the individual and pooled distributions using the following code (note that we could have used the output of the `fitdist` function from the `SHELF` package if we actually elicited quantiles from an expert). Figure 8.11 illustrates the pooled plot of survival probabilities.

```
plot_opinion1<- plot_expert_opinion(param_expert_example1[[1]],
                 weights = param_expert_example1[[1]]$wi)
ggsave("Vignette_Example 1 - Expert Opinion.png")
```

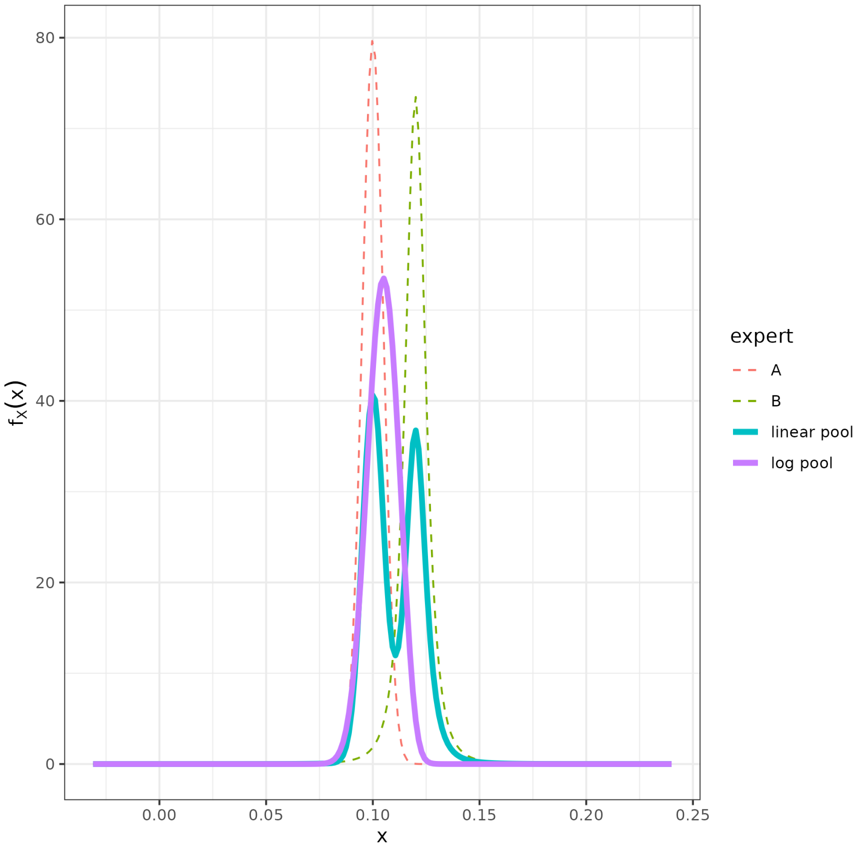


Figure 8.11: Density of pooled and individual expert opinions for survival probabilities

For the log pool we have a uni-modal distribution (in contrast to the bi-modal linear pool) which has a 95% credible interval between 9.0 – 11.9% calculated with the function below:

```
cred_int_val <- cred_int(plot_opinion1,
                val = "log pool", interval = c(0.025, 0.975))
```

We load and fit the data as follows (in this example considering just the Weibull and Gompertz models), with `pool_type = “log pool”` specifying that we want to use the logarithmic pooling (rather than default “linear pool”). We do this as we wish to compare the results to the penalized maximum likelihood estimates in the next section.

```
data2 <- data %>% rename(status = censored) %>%
                  mutate(time2 = ifelse(time > 10, 10, time),
                        status2 = ifelse(time> 10, 0, status))

#Set the opinion type to "survival"
example1  <- fit.models.expert(formula=Surv(time2,status2)~1,data=data2,
                               distr=c("wph", "gomp"),
                               method="hmc",
                               iter = 5000,
                               pool_type = "log pool",
                               opinion_type = "survival",
                               times_expert = timepoint_expert,
                               param_expert = param_expert_example1)
```

Both visual fit and model fit statistics highlight that the Weibull model is a poor fit to both the expert opinion and data (black line referring to the 95% confidence region for the experts prior belief).

```
model.fit.plot(example1, type = "dic")
#N.B. plot.expertsurv (ported directly from survHE) plots the survival
↪  function at the posterior mean parameter values
#while it is more robust to use the entire posterior sample (make.surv),
↪  however, in this case both results are similar.
plot(example1, add.km = T, t = 0:30)+
     theme_light()+
     scale_x_continuous(expand = c(0, 0), limits = c(0,NA), breaks=seq(0, 30,
     ↪  2)) +
     scale_y_continuous(expand = c(0, 0), limits = c(0, NA), breaks=seq(0, 1,
     ↪  0.05))+
     geom_segment(aes(x = 14, y = cred_int_val[1], xend = 14, yend =
     ↪  cred_int_val[2]))
```

Survival functions including expert opinion are presented in Figure 8.12.

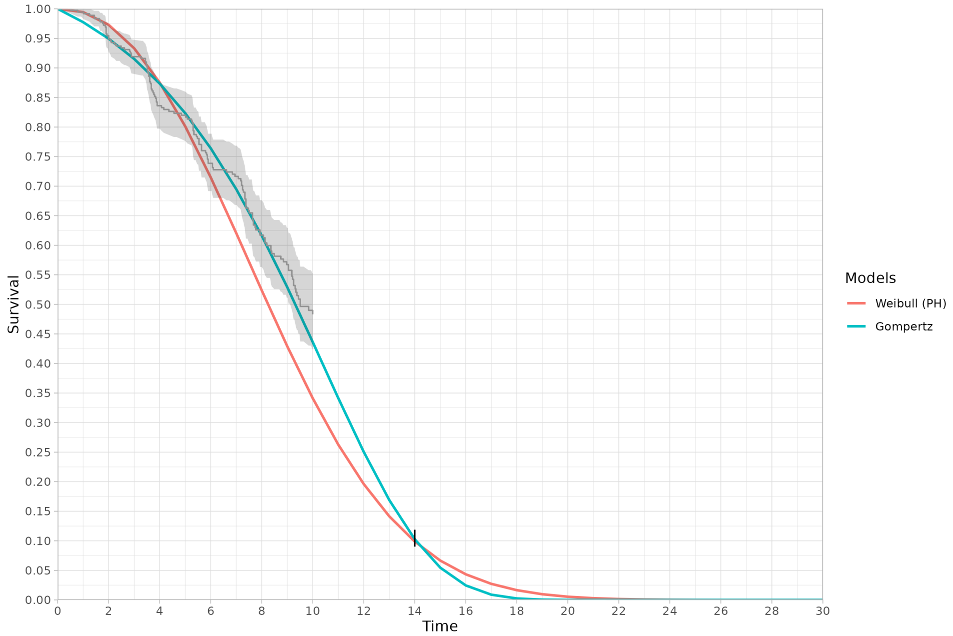


Figure 8.12: Predicted survival functions for each of the parametric models including expert opinion

The goodness of fit for each of the parameteric models including expert opinion are presented in Figure 8.13.

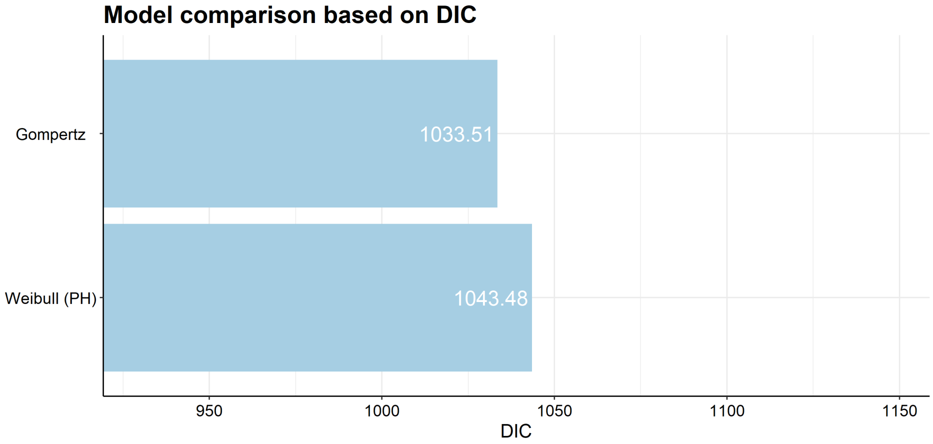


Figure 8.13: DIC values for each of the parametric survival models

8.2.3 Expert Opinion using Penalized Maximum Likelihood

We can also fit the model by Penalized Maximum Likelihood approaches through the `flexsurv` package (Jackson 2016). All that is required that the `method="hmc"` is changed to `method="mle"` with the `iter` argument now redundant. One argument that maybe of interest is the `method_mle` which is the optimization procedure that `flexsurv` uses. In case the optimization fails, we can sometimes obtain convergence with the

Nelder-Mead algorithm. If the procedure is still failing, it may relate to the expert opinion being too informative or in conflict with the observed data.

It should be noted that the results will be similar to the Bayesian approach when the expert opinion is unimodal (as maximum likelihood produces a point estimate) and relatively more informative, therefore we use the logarithmic pool which is unimodal.

We find that the AIC values also favour the Gompertz model by a large factor (not shown) and are very similar to the DIC presented for the Bayesian model.

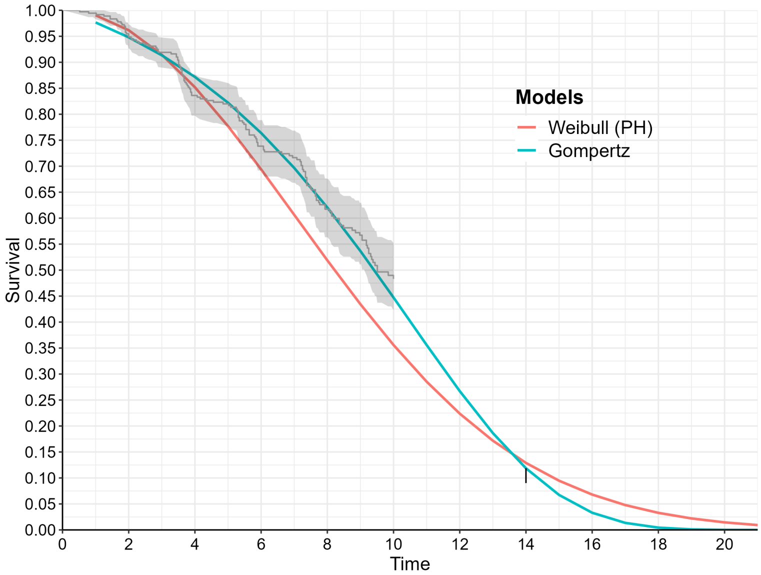


Figure 8.14: Predicted survival functions with expert opinion using penalized maximum likelihood

expertsurv modifies some of the flexsurv functions, so if you wish to use revert to the original flexsurv functions within the same session you should run the following commands:

```
unloadNamespace("flexsurv") #Unload flexsurv and associated name
↪  spaces
require("flexsurv") #reload flexsurv
```

8.2.4 Expert Opinion on Survival of a comparator arm

In this situation we place an opinion on the comparator arm.

```
    param_expert_example2[[1]] <- data.frame(dist = c("norm"),
                                             wi = c(1),
                                             param1 = c(0.1),
                                             param2 = c(0.005),
                                             param3 = c(NA))
```

```
#Check the coding of the arm variable
#Comparator is 0, which is our id_St
unique(data$arm)
#> [1] 0 1

survHE.data.model  <- fit.models.expert(formula=Surv(time2,status2)~
                        as.factor(arm),data=data2,
                        distr=c("wei"),
                        method="hmc",
                        iter = 5000,
                        opinion_type = "survival",
                        id_St = 0,
                        times_expert = timepoint_expert,
                        param_expert = param_expert_example2)
```

We can remove the impact of expert opinion by running the same model in the `survHE` package. Alternatively we note that a $\mathcal{B}(1,1)$ distribution is uniform on the survival probability and does not change the likelihood.

```
param_expert_vague <- list()
param_expert_vague[[1]] <- data.frame(dist = "beta", wi = 1, param1 = 1,
↪  param2 = 1, param2 = NA)
```

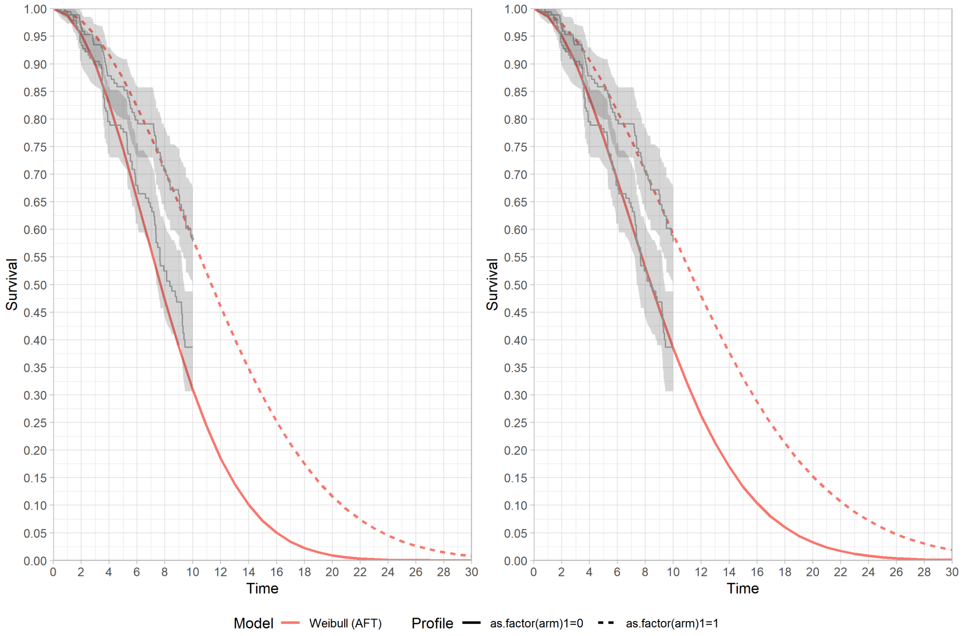


Figure 8.15: Predicted survival functions with expert information (left) and without expert information (right)

The survival function for "arm 1" has been shifted downwards slightly, however the covariate for the accelerated time factor has markedly increased to counteract the lower survival probability for the reference (arm 0).

8.2.5 Expert Opinion on Survival Difference

This example illustrates an opinion on the survival difference. For illustration we select the Gompertz distribution, noting that a negative shape parameter will lead to a proportion of subjects living forever. Clearly the mean is not defined in these cases so the code automatically constrains the shape to be positive.

```
param_expert3 <- list()

#Prior belief of 5 "months" difference in expected survival
param_expert3[[1]] <- data.frame(dist = "norm", wi = 1, param1 = 5, param2 =
↪  0.2, param3 = NA)

# Survival difference is Mean_surv[id_trt]- Mean_surv[id_comp]
survHE.data.model  <- fit.models.expert(formula=Surv(time2,status2)~
                                                as.factor(arm),data=data2,
                                                distr=c("gom"),
                                                method="hmc",
                                                iter = 5000,
                                                opinion_type = "mean",
                                                id_trt = 1,
                                                param_expert = param_expert3)
```

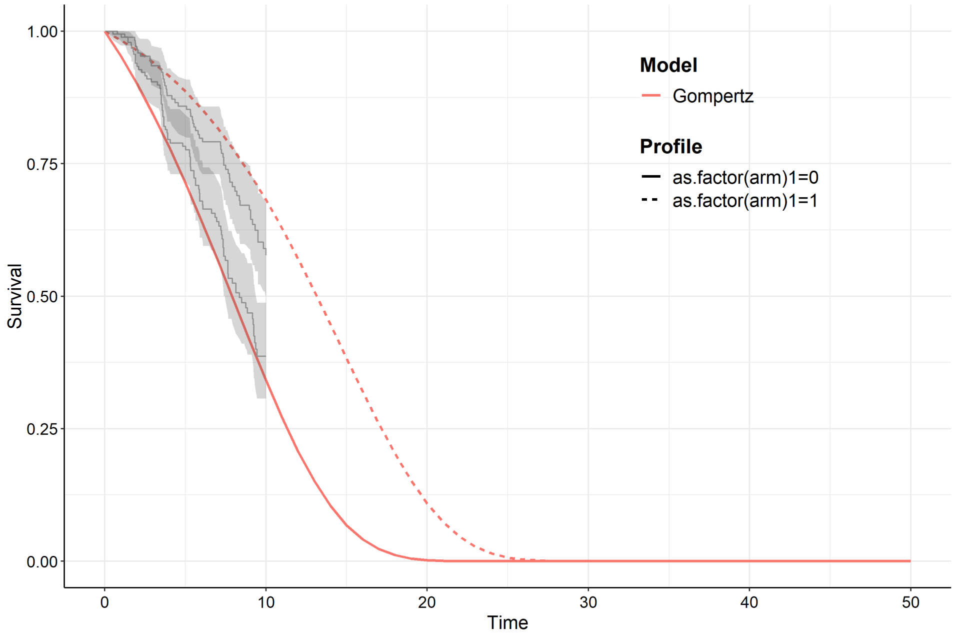


Figure 8.16: Predicted survival functions with expert information on expected differences in survival.

8.2.6 Compatibility with underlying packages survHE and flexsurv

As stated in the introduction this package relies on many of the core functions of the survHE package (Baio, 2020). Because we do not not implement expert opinion with INLA and because future versions of survHE may introduce conflicts with the current implementation, we have directly ported the key functions from survHE into the package so that expertsurv no longer imports survHE (all credit for those functions goes to Baio (2020) and co-contributors).

In theory the same concern could apply to the flexsurv package, however, this package has been released for some years and it is unlikely that the code architecture would change sufficiently to cause issues (however, for reference expertsurv was built with flexsurv = v2.0).

As mentioned, there are several modifications to flexsurv functions in order to accommodate expert opinion (by changing the functions within the namespace of the flexsurv environment). These should have no impact on the operation of flexsurv and these changes are only invoked when flexsurv is loaded. However, in the situation where you would like to revert to original flexsurv functions during the session, simply run the following:

```
    unloadNamespace("flexsurv") #Unload flexsurv and associated name spaces
    require("flexsurv") #reload flexsurv
#If this doesn't work you can use the pacman package
pacman::p_loaded("flexsurv")
```

Care should be taken, however to ensure the packages were successfully unloaded as other packages which require flexsurv can block the unloading to that package (which will cause an error).

8.2.7 Model Diagnostics

As this is a Bayesian analysis convergence diagnostics should be performed. Poor convergence can be observed for many reasons, however, because of our use of expert opinion it may be a symptom of conflict between the observed data and the expert's opinion.

Default priors should work in most situations, but still need to be considered. At a minimum the Bayesian results without expert opinion should be compared against the maximum likelihood estimates. If considerable differences are present the prior distributions should be investigated.

Because the analysis is done in JAGS and Stan we can leverage the "ggmcmc" package Fernández-i-Marín (2016):

```
#For Stan Models # Log-Normal, RP, Exponential, Weibull
ggmcmc(ggs(example1$models$`Exponential`), file = "Exponential.pdf")

#For JAGS Models # Gamma, Gompertz, Generalized Gamma
ggmcmc(ggs(as.mcmc(example1$models$`Gamma`)), file = "Gamma.pdf")
```

8.2.8 Shiny Application for `expertsurv`

R packages have many advantages, primarily allowing for easy sharing and documentation of code and ensuring code follows standardised conventions. Although R packages are more user friendly than a collection of functions, they still require the user to have a working knowledge of R something which most experts will not have.

As noted by Mikkola et al. (2023), there is a need for tools which can embed the elicitation of expert opinion within the statistical workflow. One such tool is Shiny, an open-source R package that provides an elegant and powerful web framework for building web applications using R. Shiny can turn analyses conducted by R into interactive web applications without requiring HTML, CSS, or JavaScript knowledge (Chang et al., 2023).

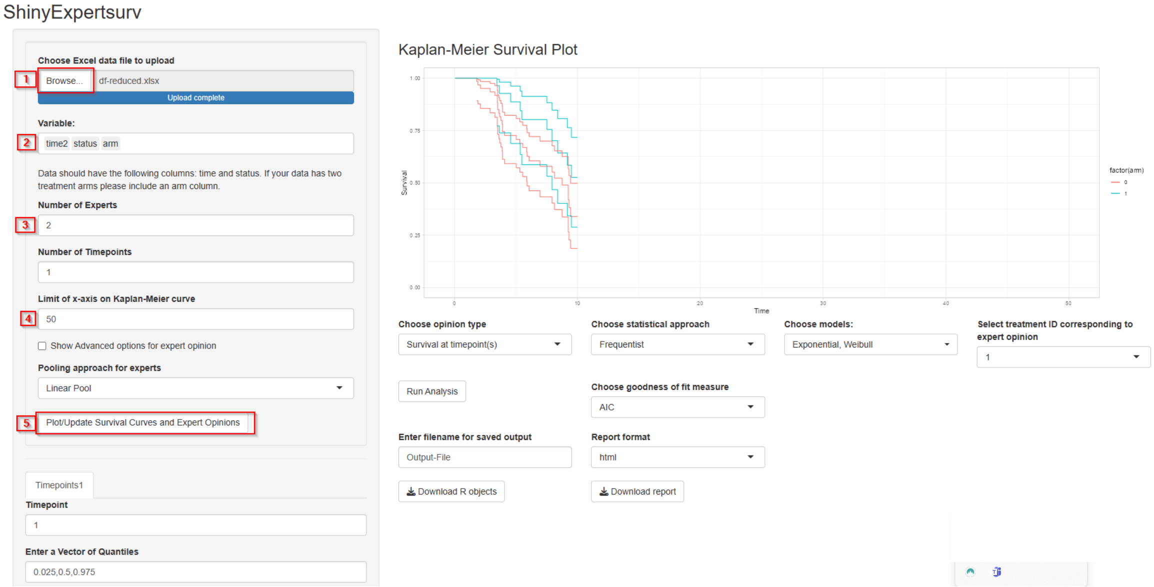
Learning how to interact with a webpage should be a much simpler task for experts and health economists not familiar with R. Furthermore, we can leverage R Markdown to create reproducible reports of the relevant outputs of the `expertsurv` package in formats such as PDF, HTML and Word.

The tutorial below provides an overview of the steps required to elicit expert opinion on the survival at a timepoint of 20 months and incorporate these beliefs with survival data. To begin we simply run the following function `elicit_surv`() which will open up a webpage.

In Figure 8.17 we have the following steps to upload the data:

1. Upload an excel file containing the survival data
2. Select the columns referring to the time, status and arm (if jointly modelling treatment and comparator)
3. Set the limit of the for the survival plot (ensuring that Choose opinion type as "Survival at timepoint(s)"
4. We assume that there are two experts who we will provide expert opinion

5. Once the above parameters are defined select the "Plot/Update Survival Curves and Expert Opinions"


ShinyExpertsurv

Choose Excel data file to upload

1 Browse... df-reduced.xlsx

Upload complete

Variable:

2 time2 status arm

Data should have the following columns: time and status. If your data has two treatment arms please include an arm column.

Number of Experts

3 2

Number of Timepoints

1

Limit of x-axis on Kaplan-Meier curve

4 50

□ Show Advanced options for expert opinion

Pooling approach for experts

Linear Pool

5 Plot/Update Survival Curves and Expert Opinions

Timepoints1

Timepoint

1

Enter a Vector of Quantiles

0.025,0.5,0.975

Kaplan-Meier Survival Plot

Choose opinion type

Survival at timepoint(s)

Choose statistical approach

Frequentist

Choose models:

Exponential, Weibull

Select treatment ID corresponding to expert opinion

1

Run Analysis

Choose goodness of fit measure

AIC

Enter filename for saved output

Output-File

Report format

html

Download R objects

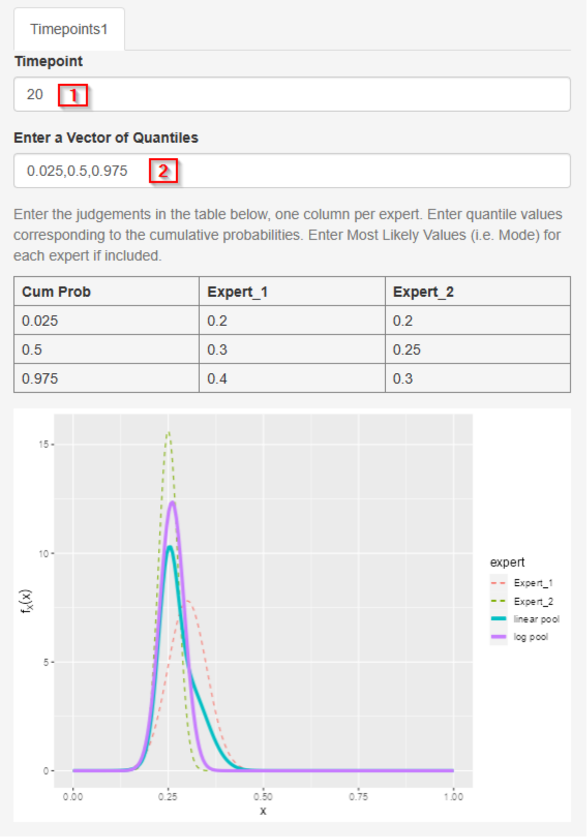
Download report

Figure 8.17: `expertsurv` Shiny application: Upload data and generate Kaplan-Meier plot

In Figure 8.18 we have the following steps to elicit the expert's opinions:

1. We set the timepoint at which we elicit expert opinion to 20 months.
2. Experts will be asked for their beliefs on the median survival at 20 months, the lower 2.5% and upper 97.5% probabilities of the population survival.

Once these steps are complete, we click the "Plot/Update Survival Curves and Expert Opinions" button.



Timepoint

20

Enter a Vector of Quantiles

0.025,0.5,0.975

Enter the judgements in the table below, one column per expert. Enter quantile values corresponding to the cumulative probabilities. Enter Most Likely Values (i.e. Mode) for each expert if included.

Cum Prob	Expert_1	Expert_2
0.025	0.2	0.2
0.5	0.3	0.25
0.975	0.4	0.3

Figure 8.18: `expertsurv` Shiny application: Include expert beliefs about survival at 20 months

In Figure 8.18 we have the following steps to run the analysis:

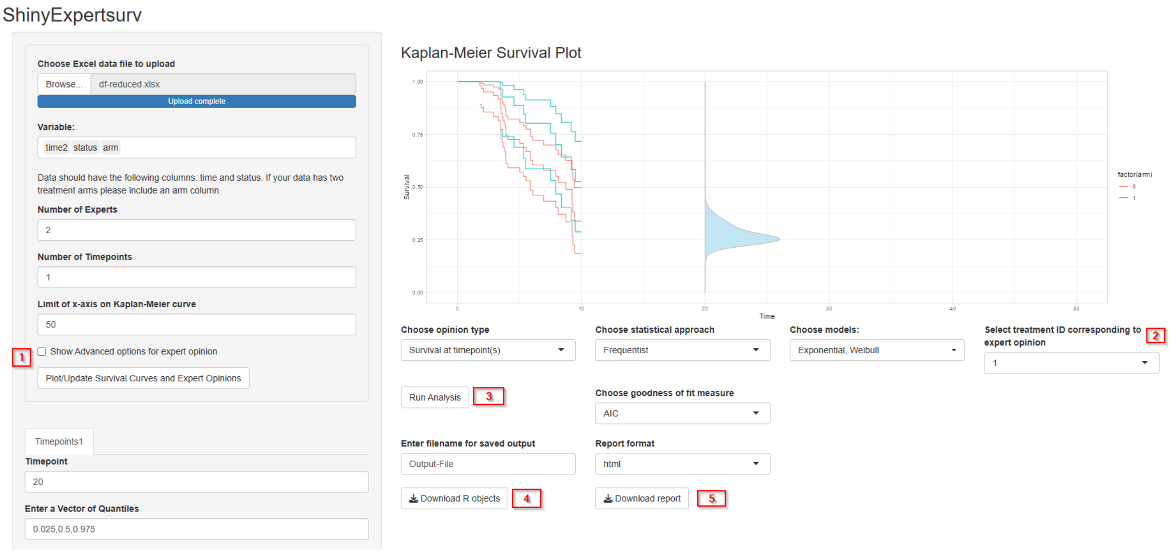
1. Several other advanced options relating to the expert opinion are available in the checkbox; these include specifying the most likely value (MLV), the choice of pooling expert opinion (linear or logarithmic pooling - default linear pooling) and the parametric distribution to fit to each individual opinion (default is best fitting).

2. We can select which treatment arm relates to the expert's belief (i.e. treatment or comparator).

3. Selecting Run Analysis will conduct the analysis; and it should be noted that the Bayesian analysis can take a considerable amount of time.

4. Once the analysis is complete we can download the expertsurv object which can be loaded an R session later.

5. We can also download the results as a report in one of three formats (HTML, PDF or Word) as shown in Figure 8.20.



ShinyExpertsurv

Choose Excel data file to upload: df-reduced.xlsx

Upload complete

Variable: time2 status arm

Data should have the following columns: time and status. If your data has two treatment arms please include an arm column.

Number of Experts: 2

Number of Timepoints: 1

Limit of x-axis on Kaplan-Meier curve: 50

1 ☐ Show Advanced options for expert opinion

Timepoints1

Timepoint: 20

Enter a Vector of Quantiles: 0.025,0.5,0.975

Choose opinion type: Survival at timepoint(s)

Choose statistical approach: Frequentist

Choose models: Exponential, Weibull

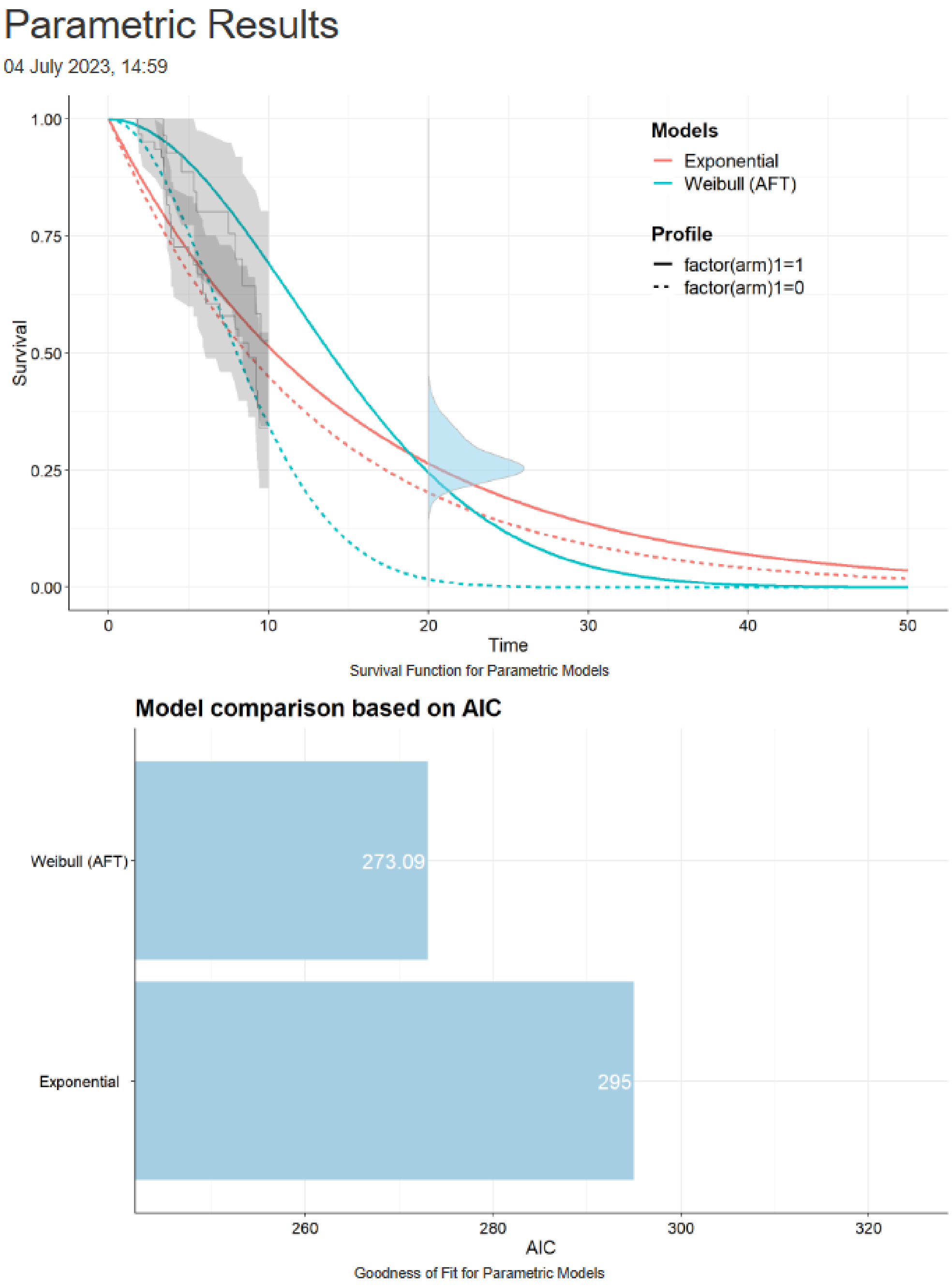
Select treatment ID corresponding to expert opinion: 1 2

Choose goodness of fit measure: AIC

Enter filename for saved output: Output-File

Report format: html

Figure 8.19: expertsurv Shiny application: Running statistical analysis



Parametric Results

04 July 2023, 14:59

Survival Function for Parametric Models

Model comparison based on AIC

Goodness of Fit for Parametric Models

Figure 8.20: `expertsurv` Shiny application: Results generated from the Markdown file

Part V

Conclusion

9 Conclusion

Over the course of this thesis we have investigated novel statistical methods for the purpose of extrapolating survival outcomes. The thesis contributions are organised into three parts: change-point modelling; incorporating expert opinion; and software modelling. The methods provide significant contributions in both theory and practice of decision modelling with time-to-event outcomes along with broader contributions relating to the inclusion of expert opinion with statistical models.

9.1 Change-point Survival Models

In the first part of the thesis we developed a novel and computationally efficient algorithm to estimate the number and location of change-points for survival models. The Bayesian approach allowed us to propagate the uncertainty in both the change-point locations and numbers, something which is not readily apparent when considering a frequentist approach.

The piecewise exponential survival models improves upon the standard practice of visually assessing timepoints at which a constant hazard is plausible and considered a fully parametric approach which accounts for the uncertainty in the location of these timepoints. The fully parametric nature of the model is important as it allows for statistical comparison with other survival models considered in decision modelling. Another contribution is the estimation of more complex parametric change-point models which include various scenarios of interest to economic modellers, such as non-constant hazards, converging hazards, and change-points in the hazard ratio between a treatment and comparator.

In summary, the change-point models replace the subjective strategy of identifying change-point locations by visual assessment with an objective and statistically coherent strategy which fully propagates uncertainty in the survival function; a key requirement in decision making.

9.2 Including Expert Opinion with Statistical Models

The second major contribution relates to the incorporation of expert opinion on observable outcomes into a broad range of statistical models and in particular survival models. The method is more flexible than existing approaches, allowing for the inclusion of a wide range of beliefs for the expert opinion (including pooled distributions) into statistical models. The range of statistical models that can be estimated with this framework is diverse, from repeated measures regression models to multivariate normal distributions. Strategies to include expert opinion on observable outcomes are important as these types of opinions are much more straightforward to elicit than opinions on model parameters.

This contribution has implication for the wider statistical community and is under review in a general Bayesian statistical journal; in contrast to the other work which has been published in journals with a HTA focus Cooney and White (2023a,c). Specifically it fulfills many of the methodological requirements specified by Mikkola et al. (2023) to enable more widespread use of expert opinion.

Complementing this research, we have also investigated the relative strength of belief of expert opinions through an intuitive measure known as effective sample size for a variety of statistical models with which we incorporate expert opinion. This is an important consideration as even when expert opinion on observable quantities can be integrated with the statistical model, it is important that these opinions are elicited in a robust manner. Effective sample size could be a useful calibration tool to ensure the final elicited belief, typically represented as a probability distribution, represents the true certainty of the expert's opinion.

9.3 Software Applications

For each of the contributions listed above we recognized that the widespread utilization of the approaches depends on the level of effort required to implement methods in applied examples. Therefore, the methods presented in this thesis are also implemented as freely available open-source R software packages (`PiecewiseChangepoint` and `expertsurv`). This allows health economists and other practitioners to quickly implement these approaches and provides the users with outputs typically required for decision analysis with time-to-event outcomes.

For users who are not familiar with the R programming language we developed a Shiny user interface which allows users to elicit expert opinion, incorporate with a survival dataset and record the relevant outputs.

9.4 Summary

A recurring theme in the contributions of this thesis i.e. estimation of change-point survival models and incorporation of expert opinion into statistical models is that we replace some of the arbitrary decisions and assumptions with objective analysis. Certain assumptions are required, for example in decision analysis and in particular Health Technology Assessments due to lack of data or because the economic models that typically inform these assessments can only be a simplified representation of the true decision process.

For example, when modelling survival data, an analyst may believe that the hazard of an event becomes approximately constant after a certain timepoint. This subjective belief may be informed by many factors, such as visualising the hazard function, clinical knowledge of the disease process or even a desire for a parsimonious modelling choice. Within this assumption there is the previously arbitrary decision of where to place this timepoint or whether it's addition improved the fit of the model. The change-point models described in this thesis allow the analyst to avoid making arbitrary decisions with respect to the location of this change-point. It also allows them to justify that the improvement in fit to the data is worth (in a statistical sense) the complexity of introducing a change-point.

With respect to the research on expert opinion, a decision maker faced with censored survival data may have subjective beliefs about the survival probabilities at landmark timepoints. Rather than try to select an arbitrary parametric model which supports that belief (in which the predicted survival is similar to the expert's at the landmark timepoint), their belief is directly integrated with the observed data allowing for the selection of the model which provides the best fit to both the observed data and expert's belief.

To summarize, the key contribution throughout this thesis is that decision modellers can specify hypotheses which can be potentially evaluated by the data rather than requiring additional choices from the analyst. For the change-point models the change-points themselves are estimated by the data and not the analyst, while in the situation where expert opinions are incorporated, the relative strengths of both the data and the opinion are coherently synthesized.

9.5 Future Research

Despite the numerous contributions there are several avenues for further research which could further improve the extrapolation of survival data. For change-point problems, although the code for estimating the piecewise exponential model has been implemented as a R-package, this could be extended to include the other change-point survival models considered in Chapter 5, similar to the mcp R package by Lindeløv (2020) which estimates general change-point models.

In terms of incorporation of expert opinion with survival models there are a number of potential additional research topics. Of particular importance is incorporating opinion at multiple timepoints, which currently treats each timepoint as an independent piece of information. Eliciting opinion for multiple timepoints as conditional survival or by using copulas could offer potential solutions. Quantifying the strength of expert opinion is an important and active area of research, with the research on effective sample size described in this thesis providing insight on the expert's opinion relative to number of observations in a dataset. Further work implementing some of the approaches to quantifying effective sample size such as Reimherr et al. (2021) or Hobbs et al. (2013) should be considered for parametric survival models as these could separate the contribution of expert's opinion from the existing data. Additionally, investigating the effective samples sizes for the range of other inputs which may be elicited as inputs in a HTA such as Hazard Ratios and Relative Risks would be a useful activity.

Research on the optimal methods and guidance for eliciting expert opinion in survival analysis (extending the general guidance provided by Bojke et al. (2021)) would be important, for example, how best to elicit survival probabilities at a timepoint. Using the method described in this thesis we are free to specify almost any functional form for the expert's belief and we have noted that for individual experts, standard distributions will not adequately represent skewness in the distribution, for example where the mode (or most likely value) is different from the median. Further research using more complex probability density functions e.g. generalized gamma or even splines to model the elicited quantiles which can modified in an interactive manner by the expert i.e. chips and pins method (Gore, 1987) would prove an valuable extension to the expertsurv package and associated Shiny application.

Bibliography

A. E. Gelfand, D. K. D. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Achcar, J. A. and Bolfarine, H. (1989). Constant hazard against a change-point alternative: a Bayesian approach with censored data. *Communications in Statistics - Theory and Methods*, 18(10):3801–3819.

Achcar, J. A. and Loibel, S. (1998). Constant Hazard Function Models with a Change Point: A Bayesian Analysis Using Markov Chain Monte Carlo Methods. *Biometrical Journal*, 40(5):543–555.

Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle. In H Akaike, Selected Papers of Hirotugu Akaike*, pages 199–213. Springer New York.

Al-Awadhi, S. A. and Garthwaite, P. H. (1998). An elicitation method for multivariate normal distributions. *Communications in Statistics - Theory and Methods*, 27(5):1123–1142.

Anis, M. Z. (2009). Inference on a Sharp Jump in Hazard Rate: A Review. *Stochastics and Quality Control*, 24(2):213–229.

Ayers, D., Cope, S., Towle, K., Mojebi, A., Marshall, T., and Dhanda, D. (2022). Structured expert elicitation to inform long-term survival extrapolations using alternative parametric distributions: a case study of CAR T therapy for relapsed/ refractory multiple myeloma. *BMC Medical Research Methodology*, 22.

Bagust, A. and Beale, S. (2014). Survival Analysis and Extrapolation Modeling of Time-to-Event Clinical Trial Data for Economic Evaluation: An Alternative Approach. *Medical Decision Making*, 34(3):343–351.

Baio, G. (2020). survHE: Survival Analysis for Health Economic Evaluation and Cost-Effectiveness Modeling. *Journal of Statistical Software*, 95(14):1–47.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association*, 91(436):1450–1460.

Bell Gorrod, H., Kearns, B., Stevens, J., and 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.

Bernardo, J. and Smith, A. (2000). *Bayesian Theory* (2nd ed.). Wiley.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):1103–1130.

Bojke, L., Soares, M., Claxton, K., Colson, A., Fox, A., Jackson, C., Jankovic, D., Morton, A., Sharples, L., and Taylor, A. (2021). Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. 25:37.

Bølstad, J. (2019). *How Efficient is Stan Compared to JAGS? Conjugacy, Pooling, Centering, and Posterior Correlations*. http://www.boelstad.net/post/stan_vs_jags_speed.

Bousquet, N. (2006). *A Bayesian analysis of industrial lifetime data with Weibull distributions*. INRIA. https://inria.hal.science/inria-00115528v4/document.

Briggs, A., Sculpher, M., and Claxton, K. (2006). *Decision Modelling for Health Economic Evaluation*. Oxford University Press.

Briggs, A. and Tambour, M. (1998). The design and analysis of stochastic cost-effectiveness studies for the evaluation of health care interventions. *Therapeutic Innovation and Regulatory Science*, 35.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Bullement, A., Latimer, N., and Bell Gorrod, H. (2019). Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study. *Value in Health*, 22(3):276–283.

Bürkner, P.-C. (2017). brms: A R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1):1–28.

Campodonico, S. and Singpurwalla, N. (1993). *Expert Opinion in Reliability*. Defense Technical Information Center. https://apps.dtic.mil/sti/pdfs/ADA293921.pdf.

Canadian Agency for Drugs and Technologies in Health (2017). *Guidelines for the*

Economic Evaluation of Health Technologies: Canada. https://www.cadth.ca/guidelines-economic-evaluation-health-technologies-canada-0.

Castelloe, J. M. and Zimmerman, D. L. (2002). *Convergence Assessment for Reversible Jump MCMC*. University of Iowa. http://www.jmcastelloe.net/pubs/JMCConvDiagTr313.pdf.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). *shiny: Web Application Framework for R.* R package version 1.7.4.1. https://shiny.posit.co/.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., et al. (2011). Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *New England Journal of Medicine*, 364(26):2507–2516.

Chapple, A., Peak, T., and Hemal, A. (2020). A novel Bayesian continuous piecewise linear log-hazard model, with estimation and inference via reversible jump Markov chain Monte Carlo. *Statistics in Medicine*, 39.

Che, Z., Green, N., and Baio, G. (2023). Blended Survival Curves: A New Approach to Extrapolation for Time-to-Event Outcomes from Clinical Trials in Health Technology Assessment. *Medical Decision Making*, 43(3):299–310.

Clemen, R. T. and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19:187–203.

ClinicalTrials.gov. (2023). *Database of privately and publicly funded clinical studies conducted around the world*. https://clinicaltrials.gov/ct2/home.

Cole, S., Chu, H., and Greenland, S. (2013). Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology*, 179(2):252–260.

Collett, D. (2015). *Modelling Survival Data in Medical Research* (3rd ed). CRC Press.

Coolen, F. (1996). On Bayesian reliability analysis with informative priors and censoring. *Reliability Engineering & System Safety*, 53(1):91–98.

Cooney, P. and White, A. (2023a). Direct Incorporation of Expert Opinion into Parametric Survival Models to Inform Survival Extrapolation. *Medical Decision Making*, 43(3):325–336.

Cooney, P. and White, A. (2023b). *expertsurv: Incorporate Expert Opinion with Parametric Survival Models.* R package version 1.1.0. https://cran.irsn.fr/web/packages/expertsurv/index.html.

Cooney, P. and White, A. (2023c). Extending Beyond Bagust and Beale: Fully Parametric

Piecewise Exponential Models for Extrapolation of Survival Outcomes in Health Technology Assessment. *Value in Health*, 26(10):1510–1517.

Cooper, M., Smith, S., Williams, T., and Aguiar-Ibáñez, R. (2022). How accurate are the longer-term projections of overall survival for cancer immunotherapy for standard versus more flexible parametric extrapolation methods? *J Med Econ*, 25(1):260–273.

Cope, S., Ayers, D., Zhang, J., Batt, K., and Jansen, J. P. (2019). Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia. *BMC Med Res Methodol*, 19(1):182.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M., and Tenenbaum, D. (2023). *remotes: R Package Installation from Remote Repositories, Including 'GitHub'*. R package version 2.4.2. https://github.com/r-lib/remotes.

Daneshkhah, A. and Oakley, J. (2010). Eliciting multivariate probability distributions. In K Böcker*, Rethinking risk measurement and reporting*, volume 1, page 23. Risk Books.

Davies, C., Briggs, A., Lorgelly, P., Garellick, G., and Malchau, H. (2013). The "Hazards" of Extrapolating Survival Curves. *Medical Decision Making*, 33(3):369–380.

Dias, S., Welton, N. J., Sutton, A. J., and AE, A. (2011). *NICE DSU TECHNICAL SUPPORT DOCUMENT 5:Evidence synthesis in the baseline natural history model*. https://www.sheffield.ac.uk/nice-dsu/tsds/evidence-synthesis/.

Drummond, M., Sculpher, M., Claxton, K., Stoddart, G., and Torrance, G. (2015). *Methods for the Economic Evaluation of Health Care Programmes.* (4th ed.). Oxford University Press.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Egger, M., May, M., Chêne, G., Phillips, A. N., et al. (2002). Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *The Lancet*, 360(9327):119–129.

Estève, J., Benhamou, E., Croasdale, M., and Raymond, L. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, 9(5):529–538.

Fackler, P. L. (1991). Modeling Interdependence: An Approach to Simulation and Elicitation. *American Journal of Agricultural Economics*, 73(4):1091–1097.

Fang, L. and Su, Z. (2011). A hybrid approach to predicting events in clinical trials with time-to-event outcomes. *Contemporary Clinical Trials*, 32(5):755–759.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

Fernández-i-Marín, X. (2016). ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, 70(9):1–20.

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4):507–521.

Fitzpatrick, P. (2009). *Advanced Calculus*. American Mathematical Society, second edition.

Fleurence, R. and Hollenbeak, C. (2007). Rates and probabilities in economic modelling: Transformation, translation and appropriate application. *PharmacoEconomics*, 25(1):3–6.

Garthwaite, P. H. and Al-Awadhi, S. A. (2001). Non-Conjugate Prior Distribution Assessment for Multivariate Normal Sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):95–110.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*, 100(470):680–701.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, 7(4):457–472.

Geweke, J. (1991). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. Federal Reserve Bank of Minneapolis. https://minneapolisfed.org/Research/sr/sr148.pdf.

Ghosh, J. K., Joshi, S. N., and Mukhopadhyay, O. (1996). Asymptotics of a Bayesian approach to estimating change-point in a hazard rate. *Communications in Statistics - Theory and Methods*, 25(12):3147–3166.

Ghosh, S. K. and Ebrahimi, N. (2008). Bayesian Analysis of Change-Point Hazard Rate Problem. *Journal of Statistical Theory and Practice*, 2(4):523–533.

Gierz, K. and Park, K. (2022). Detection of multiple change points in a Weibull accelerated failure time model using sequential testing. *Biometrical Journal*, 64(3):617–634.

Gijbels, I. and Gurler, U. (2003). Estimation of a Change Point in a Hazard Function Based on Censored Data. *Lifetime Data Analysis*, 9(4):395–411.

Glucksberg, H., Cheever, M. A., Farewell, V. T., Fefer, A., Sale, G. E., and Thomas, E. D. (1981). High-dose combination chemotherapy for acute nonlymphoblastic leukemia in adults. *Cancer*, 48(5):1073–1081.

Goodman, M. S., Li, Y., and Tiwari, R. C. (2011). Detecting multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics*, 38(11):2523–2532.

Gore, S. M. (1987). Biostatistics and the medical research council. *Medical Research Council News*, 35:19–20.

Gorrod, H. B., Kearns, B., Stevens, J., Thokala, P., Labeit, A., Latimer, N., Tyas, D., and Sowdani, A. (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.

Grupp, S., Laetsch, T., Buechner, J., Bittencourt, H., Maude, S., et al. (2016). Analysis of a Global Registration Trial of the Efficacy and Safety of CTL019 in Pediatric and Young Adults with Relapsed/Refractory Acute Lymphoblastic Leukemia (ALL). *Blood*, 128:221–221.

Grupp, S. A., Maude, S. L., Rives, S., Baruchel, A., Boyer, M. W., et al. (2018). Updated Analysis of the Efficacy and Safety of Tisagenlecleucel in Pediatric and Young Adult Patients with Relapsed/Refractory (r/r) Acute Lymphoblastic Leukemia. *Blood*, 132(Supplement 1):895–895.

Grupp, S. A., Maude, S. L., Shaw, P. A., et al. (2015). Durable Remissions in Children with Relapsed/Refractory ALL Treated with T Cells Engineered with a CD19-Targeted Chimeric Antigen Receptor (CTL019). *Blood*, 126(23):681.

Guyot, P., Ades, A. E., Beasley, M., Lueza, B., Pignon, J.-P., and Welton, N. J. (2017). Extrapolation of Survival Curves from Cancer Trials Using External Information. *Medical Decision Making*, 37(4):353–366.

Guyot, P., Ades, A. E., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, 12(1):9.

Hagar, Y. and Dukic, V. (2015). *Comparison of hazard rate estimation in R.* arXiv. `https://arxiv.org/abs/1509.03253`.

Hartmann, M., Agiashvili, G., Bürkner, P., and Klami, A. (2020). *Flexible Prior Elicitation via the Prior Predictive Distribution*. arXiv. `https://arxiv.org/abs/2002.09868`.

Haute Autorité de Santé (2020). *Choices in Methods for Economic Evaluation*. https://www.has-sante.fr/jcms/r_1499251/en/choices-in-methods-for-economic-evaluation.

Health Information and Quality Authority (2020). *Guidelines for the Economic Evaluation of Health Technologies in Ireland*. https://www.hiqa.ie/sites/default/files/2020-09/HTA-Economic-Guidelines-2020.pdf.

Herbst, R. S., Garon, E. B., et al. (2021). Five Year Survival Update From KEYNOTE-010: Pembrolizumab Versus Docetaxel for Previously Treated, Programmed Death-Ligand 1-Positive Advanced NSCLC. *Journal of Thoracic Oncology*, 16(10):1718–1732.

Hess, K. and Gentleman, R. (2019). *muhaz: Hazard Function Estimation in Survival Analysis*. R package version 1.2.6.1. https://CRAN.R-project.org/package=muhaz.

Hobbs, B., Carlin, B., and Sargent, D. (2013). Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10:430–440.

Hosack, G. R., Hayes, K. R., and Barry, S. C. (2017). Prior elicitation for Bayesian generalised linear models with application to risk control option assessment. *Reliability Engineering & System Safety*, 167:351–361.

Ibrahim, J., Sinha, J., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Stat. Med.*, 34(28):3724–3749.

Jackson, C. (2016). flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software*, 70(8):1–33.

Jackson, C. (2023). *survextrap: a package for flexible and transparent survival extrapolation*. arXiv. https://arxiv.org/abs/2306.03957.

Jackson, C., Stevens, J., Ren, S., Latimer, N., Bojke, L., Manca, A., and Sharples, L. (2017). Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods. *Medical Decision Making*, 37(4):377–390.

Jackson, C. H., Sharples, L. D., and Thompson, S. G. (2010). Structural and parameter uncertainty in Bayesian cost-effectiveness models. *J R Stat Soc Ser C Appl Stat*, 59(2):233–253.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford.

Jewell, N. P. (1982). Mixtures of Exponential Distributions. *The Annals of Statistics*, 10(2):479–484.

Johnson, W. O. (1996). Predictive Influence in the Log Normal Survival Model. In J. C. Lee et al. (eds.) *Modelling and Prediction Honoring Seymour Geisser*, pages 104–121.

Kadane, J. and Wolfson, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association*, 75(372):845–854.

Kamgar, F., Ho, S., Hawe, E., and Brodtkorb, T. H. (2022). *A Review of Treatment Effect Waning Methods for Immuno-Oncology Therapies in National Institute for Health and Care Excellence Technology Appraisals*. [Poster Presentation]. ISPOR Europe 2022. doi: 10.1016/j.jval.2022.09.476.

Karasoy, D. S. and Kadilar, C. (2007). A new Bayes estimate of the change point in the hazard function. *Computational Statistics and Data Analysis*, 51(6):2993–3001.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kearns, B., Stevenson, M. D., Triantafyllopoulos, K., and Manca, A. (2019). Generalized Linear Models for Flexible Parametric Modeling of the Hazard Function. *Medical Decision Making*, 39(7):867–878.

Kepner, J. L., Keith, S. Z., and Harper, J. D. (1989). A Note on Evaluating a Certain Orthant Probability. *The American Statistician*, 43(1):48–49.

Kim, J., Cheon, S., and Jin, Z. (2020). Bayesian multiple change-points estimation for hazard with censored survival data from exponential distributions. *Journal of the Korean Statistical Society*, 49(1):15–31.

Klienbaum, M. and Klein, D. G. (2016). *Survival Analysis : A self-learning text,* (3rd ed.). Springer-Verlag.

Klijn, S. L., Fenwick, E., Kroep, S., Johannesen, K., Malcolm, B., Kurt, M., Kiff, C., and Borrill, J. (2021). What Did Time Tell Us? A Comparison and Retrospective Validation of Different Survival Extrapolation Methods for Immuno-Oncologic Therapy in Advanced or Metastatic Renal Cell Carcinoma. *Pharmacoeconomics*, 39(3):345–356.

Kosinski, M. and Biecek, P. (2020). *RTCGA: The Cancer Genome Atlas Data Integration*. R package version 1.18.0. `https://www.bioconductor.org/packages/release/bioc/html/RTCGA.html`.

Latimer, N. R. (2013). *NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data.* `http://nicedsu.org.uk/`.

Latimer, N. R. (2014). Response to "Survival Analysis and Extrapolation Modeling of Time-to-Event Clinical Trial Data for Economic Evaluation: An Alternative Approach" by Bagust and Beale. *Medical Decision Making*, 34(3):279–282.

Lele, S. and Das, A. (2000). Elicited Data and Incorporation of Expert Opinion for Statistical Inference in Spatial Studies. *Mathematical Geology*, 32:465–487.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Lin, S.-W. and Bier, V. (2008). A study of expert overconfidence. *Reliability Engineering & System Safety*, 93:711–721.

Lindeløv, J. K. (2020). *mcp: A R Package for Regression With Multiple Change Points* OSF Preprints. *OSF Preprints*. `https://osf.io/fzqxv`.

Littell, R. C. (1990). *Analysis of Repeated Measures Data*. [Paper presentation]. 2nd Annual Conference on Applied Statistics in Agriculture.

Low Choy, S., Murray, J., James, A., Mengersen, K. L., and West, M. (2013). Indirect elicitation from ecological experts: From methods and software to habitat modelling and rock-wallabies. In *The Oxford Handbook of Applied Bayesian Analysis*. Oxford University Press.

Mandel, M. (2013). Simulation-Based Confidence Intervals for Functions With Complicated Derivatives. *The American Statistician*, 67(2):76–81.

Matthews, D. E. and Farewell, V. T. (1982). On Testing for a Constant Hazard against a Change-Point Alternative. *Biometrics*, 38(2):463–468.

Matthews, D. E. and Farewell, V. T. (1985). On a singularity in the likelihood for a change-point hazard rate model. *Biometrika*, 72(3):703–704.

Matthews, D. E., Farewell, V. T., and Pyke, R. (1985). Asymptotic Score-Statistic Processes and Tests for Constant Hazard Against a Change-Point Alternative. *Ann. Statist.*, 13(2):583–591.

Maude, S. L., Pulsipher, M. A., Boyer, M. W., Grupp, S. A., Davies, S. M., et al. (2016). Efficacy and Safety of CTL019 in the First US Phase II Multicenter Trial in Pediatric Relapsed/Refractory Acute Lymphoblastic Leukemia: Results of an Interim Analysis. *Blood*, 128(22):2801.

McNulty, G. (2021). *The Pareto-Gamma Mixture*. Casualty Actuarial Society. https://www.casact.org/sites/default/files/2021-05/McNulty_The_Pareto_Gamma_Mixture.pdf.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2023). Prior Knowledge Elicitation: The Past, Present, and Future. *Bayesian Analysis*, pages 1 – 33.

Miller, R. and Halpern, J. (1982). Regression with Censored Data. *Biometrika*, 69(3):521–531.

Monnickendam, G., Zhu, M., McKendrick, J., and Su, Y. (2019). Measuring Survival Benefit in Health Technology Assessment in the Presence of Nonproportional Hazards. *Value in Health*, 22(4):431–438.

Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.

Müller, H. G. and Wang, J.-L. (1994). Change-Point Models for Hazard Functions. *Lecture Notes-Monograph Series*, 23:224–241.

National Institute for Health and Care Excellence (2022). *NICE health technology evaluations: the manual*. https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705 – 767.

Neuenschwander, B., Weber, S., Schmidli, H., and O'Hagan, A. (2020). Predictively consistent prior effective sample sizes. *Biometrics*, 76(2):578–587.

Nguyen, H. T., Rogers, G. S., and Walker, E. A. (1984). Estimation in Change-Point Hazard Rate Models. *Biometrika*, 71(2):299–304.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization* (2nd ed.). Springer Science & Business Media.

Oakley, J. (2020). *SHELF: Tools to Support the Sheffield Elicitation Framework*. R package version 1.7.0. https://CRAN.R-project.org/package=SHELF.

Office for National Statistics (ONS) (2021). *Dataset: Mortality rates (qx), by single year of age*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/mortalityratesqxbysingleyearofage.

O'Hagan, A. (2008). Chapter 6 The Bayesian Approach to Statistics.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd.

Ouwens, M. (2018). *Use of clinical opinion in the estimation of survival extrapolation distributions*. ISPOR EU. https://www.ispor.org/docs/default-source/presentations/91714pdf.pdf?sfvrsn=a5b2756f_0.

O'Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, 73(sup1):69–81.

Palmeros, O., Villaseñor, J. A., and González, E. (2018). On computing estimates of a change-point in the Weibull regression hazard model. *Journal of Applied Statistics*, 45(4):642–648.

Percy, D. F. (2004). Subjective priors for maintenance models. *Journal of Quality in Maintenance Engineering*, 10:221–227.

Pharmaceutical Benefits Advisory Committee (2016). *Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (PBAC), Version 5.0*. https://pbac.pbs.gov.au/.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling* [Paper presentation]. DSC 2003 Working Papers. https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf.

Plummer, M. (2022). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-14. https://CRAN.R-project.org/package=rjags.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. (1986). Choosing Models for Cross-Classifications. *American Sociological Review*, 51:145.

Raftery, A. E. and Lewis, S. (1992). How Many Iterations in the Gibbs Sampler? In *Bayesian Statistics 4*, pages 763–773. Oxford University Press.

R Development Core Team (2023). R: A language and environment for statistical computing. https://www.R-project.org/.

Rebora, P., Salim, A., and Reilly, M. (2018). *bshazard: Nonparametric Smoothing of the Hazard Function*. R package version 1.1. https://cran.r-project.org/web/packages/bshazard/index.html.

Reimherr, M., Meng, X., and Nicolae, D. L. (2021). Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(3).

Robert, C., Grob, J. J., et al. (2019). Five-Year Outcomes with Dabrafenib plus Trametinib in Metastatic Melanoma. *New England Journal of Medicine*, 381(7):626–636.

Rohatgi, A. (2022). *WebPlotDigitizer: Version 4.6. Web based tool to extract data from plots images and maps.* `https://automeris.io/WebPlotDigitizer/`.

Royston, P. and Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197.

Royston, P. and Parmar, M. K. B. (2013). Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(1):152.

Roze, S., Bertrand, N., Eberst, L., and Borget, I. (2023). Projecting overall survival in health-economic models: uncertainty and maturity of data. *Curr Med Res Opin*, 39(3):367–374.

Rutherford, M. J., Lambert, P. C., Sweeting, M. J., Pennington, B., Crowther, M. J., Abrams, K. R., and Latimer, N. R. (2020). *NICE DSU TECHNICAL SUPPORT DOCUMENT 21: Flexible Methods for Survival Analysis*. `http://nicedsu.org.uk/flexible-methods-for-survival-analysis-tsd/`.

Salvo, F. D. (2008). A characterization of the distribution of a weighted sum of Gamma variables through multiple hypergeometric functions. *Integral Transforms and Special Functions*, 19(8):563–575.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

Sculpher, M. J., Claxton, K., Drummond, M., and McCabe, C. (2006). Whither trial-based economic evaluation for health care decision making? *Health economics*, 15(7):677–687.

Siegrist, K. (2023). *The Gompertz Distribution*. `http://www.randomservices.org/random/special/Gompertz.html`.

Singpurewalla, N. and Song, M. (1988). Reliability analysis using Weibull lifetime data and expert opinion. *IEEE Transactions on Reliability*, 37(3):340–347.

Sonderegger, D. (2022). *SiZer: Significant Zero Crossings*. R package version 0.1-8.https://cran.r-project.org/web/packages/SiZer/index.html.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stacy, E. W. (1962). A Generalization of the Gamma Distribution. *The Annals of Mathematical Statistics*, 33(3):1187 – 1192.

Stan Development Team (2020). *RStan: the R interface to Stan*. R package version 2.21.2. http://mc-stan.org/.

Stevens, A. J. and Longson, C. (2013). At the Center of Health Care Policy Making: The Use of Health Technology Assessment at NICE. *Medical Decision Making*, 33(3):320–324.

Surveillance, Epidemiology, and End Results (SEER) Program (2023). *Incidence - SEER Research Data*. https://seer.cancer.gov/.

TA268 (2012). National Institute for Health and Care Excellence. *Ipilimumab for previously treated advanced (unresectable or metastatic) melanoma. Technology Appraisal Guidance TA268*. https://www.nice.org.uk/guidance/ta268.

TA269 (2012). National Institute for Health and Care Excellence. *Vemurafenib for treating locally advanced or metastatic BRAF V600 mutation-positive malignant melanoma. Technology Appraisal Guidance TA269*. https://www.nice.org.uk/guidance/ta269.

TA347 (2015). National Institute for Health and Care Excellence. *Nintedanib for previously treated locally advanced, metastatic, or locally recurrent non-small-cell lung cancer. Technology Appraisal Guidance TA347*. https://www.nice.org.uk/guidance/ta347.

TA396 (2016). National Institute for Health and Care Excellence. *Trametinib in combination with dabrafenib for treating unresectable or metastatic melanoma. Technology Appraisal Guidance TA396*. https://www.nice.org.uk/Guidance/TA396.

TA428 (2017). National Institute for Health and Care Excellence. *Pembrolizumab for treating PD-L1-positive non-small-cell lung cancer after chemotherapy. Technology Appraisal Guidance TA428*. https://www.nice.org.uk/guidance/ta428.

TA447 (2017). National Institute for Health and Care Excellence. *Pembrolizumab for* *untreated PD-L1-positive metastatic non-small-cell lung cancer. Technology Appraisal Guidance TA447.* Superseded by TA531.

TA531 (2018). National Institute for Health and Care Excellence. *Pembrolizumab for untreated PD-L1-positive metastatic non-small-cell lung cancer. Technology Appraisal Guidance TA531 (Supersedes TA447).* `https://www.nice.org.uk/guidance/ta531`.

TA589 (2019). National Institute for Health and Care Excellence. *Blinatumomab for treating acute lymphoblastic leukaemia in remission with minimal residual disease activity. Technology Appraisal Guidance TA589*. `https://www.nice.org.uk/guidance/ta589`.

van Oostrum, I., Ouwens, M., Remiro-Azócar, A., Baio, G., Postma, M. J., Buskens, E., and Heeg, B. (2021). Comparison of Parametric Survival Extrapolation Approaches Incorporating General Population Mortality for Adequate Health Technology Assessment of New Oncology Drugs. *Value in Health*, 24(9):1294–1301.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.4.1. `https://mc-stan.org/loo/`.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian Information Criterion for Censored Survival Models. *Biometrics*, 56(1):256–262.

Wabersich, D. and Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, 46(1):15–28.

Wang, M., Dignam, J. J., Won, M., Curran, W., Mehta, M., and Gilbert, M. R. (2015). Variation over time and interdependence between disease progression and death among patients with glioblastoma on RTOG 0525. *Neuro-Oncology*, 17(7):999–1006.

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.*, 11:3571–3594.

Weinstein, M. C. and Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *The New England journal of medicine*, 296 13:716–21.

Weisstein, E. W. (2021). MathWorld - Exponential integral. `https://mathworld.wolfram.com/ExponentialIntegral.html`.

Wesner, J. S. and Pomeranz, J. P. F. (2021). Choosing priors in Bayesian ecological models by simulating from the prior predictive distribution. *Ecosphere*, 12(9):e03739.

Wickham, H. (2019). *Advanced R,* (2 ed.). Chapman & Hall/CRC The R Series.

Williggers, Bart and Ouwens, Mario and Briggs, Andrew and Heerspink, Hiddo and Pollock, Carol and Pecoits-Filho, Roberto and Tangri, Navdeep and Kovesdy, Csaba and Wheeler, David and Garcia-Sanchez, Juan Jose (2023). The role of expert opinion in projecting long-term survival outcomes beyond the horizon of a clinical trial. *Advances in therapy*, 40.

Wongnak, P., Bord, S., Donnet, S., Hoch, T., Beugnet, F., and Chalvet-Monfray, K. (2022). A hierarchical Bayesian approach for incorporating expert opinions into parametric survival models: A case study of female Ixodes ricinus ticks exposed to various temperature and relative humidity conditions. *Ecological Modelling*, 464:109821.

Worsley, K. J. (1988). Exact Percentage Points of the Likelihood-Ratio Test for a Change-Point Hazard-Rate Model. *Biometrics*, 44(1):259–263.

Wyse, J. and Friel, N. (2010). *Simulation-based Bayesian analysis for multiple changepoints*. arXiv. `https://arxiv.org/abs/1011.2932`.

Yao, Y.-C. (1986). Maximum likelihood estimation in hazard rate models with a change-point. *Communications in Statistics - Theory and Methods*, 15(8):2455–2466.

Yao, Y.-C. (1987). A note on testing for constant hazard against a change-point alternative. *Annals of the Institute of Statistical Mathematics*, 39(2):377–383.

Yu-Sung, S. and Masanao, Y. (2015). *R2jags: Using R to Run 'JAGS'*. R package version 0.5-7. `https://CRAN.R-project.org/package=R2jags`.

Part VI

Appendix

A Piecewise Exponential Models

A.1 Marginal Likelihood for exponential survival times

The marginal likelihood is sometimes known as the probability of the data $\pi(\mathbf{y})$ associated with a statistical model and appears as the denominator in Bayes Formula

$$\pi(\mathbf{y}) = \int_\theta \pi(\mathbf{y}|\theta)\pi(\theta)d\theta,$$

where θ is the general notation for the model parameter(s). For the exponential distribution, the model parameter $\theta = \lambda$ and the probability of the data are conditional on the hyperparameters α and β for the λ parameter which we denote together as γ. Therefore the expression becomes

$$\pi(\mathbf{y}|\gamma) = \int_\theta \pi(\mathbf{y}|\theta)\pi(\theta|\gamma)d\theta.$$

The easiest way to evaluate this integral is indirectly through Bayes formula. Bayes formula is as follows:

$$\pi(\theta|\mathbf{y},\gamma) = \frac{\pi(\mathbf{y}|\theta)\pi(\theta)\pi(\gamma)}{\int_\theta \pi(\mathbf{y}|\theta)\pi(\theta|\gamma)d\theta}.$$

The conjugate prior for an exponential likelihood is the gamma distribution. Therefore given hyperparameters α and β, the posterior is a $\mathcal{G}(\alpha + D, \beta + T)$ where D is the number of events and T is the exposure time within that interval. Letting $\alpha^* = \alpha + D$ and $\beta^* = \beta + T$ and rearranging Bayes formula, it immediately follows that the marginal likelihood is the ratio of the prior normalizing factor divided by the posterior normalizing factor;

$$\int_\theta \pi(\mathbf{y}|\theta)\pi(\theta|\gamma)d\theta = \frac{\pi(\mathbf{y}|\theta)\pi(\theta)\pi(\gamma)}{\pi(\theta|\mathbf{y})} = \frac{\beta^\alpha/\Gamma(\alpha)}{(\beta^{*\alpha^*})/\Gamma(\alpha^*)}.$$

Given k change-points, we have $k+1$ segments of data and the joint marginal likelihood is the product of these $k + 1$ segments.

A.1.1 Incorporating uncertainty in hyperparameters

The marginal likelihood can be sensitive to the hyperparameters α and β and therefore can influence the posterior distribution of the change-points. To account for this uncertainty and improve the robustness of the results we can introduce a hyperprior on β. In an extra sampling step, the hazards $\lambda_{1:k+1}$ can be "uncollapsed" and sampled at each iteration. We place a hyperprior on $\beta : \beta \sim \mathcal{G}(\xi, \delta)$, with

$$\pi(\beta|\xi, \delta) = \frac{\delta^{\xi}}{\Gamma(\xi)} \beta^{\xi-1} \exp\left(-\delta\beta\right).$$

Simplifying Equation 3.4 we note that the posterior density of the change-point number and locations is proportional to the likelihood, the prior on the hazards and the hyperprior on β.

$$\begin{aligned}
\pi(k, s_1, \dots, s_k, \beta|y_{1:d}, \lambda_{1:k+1}, \alpha, \xi, \delta) &\propto \pi\left(y_{1:d}|s_1, \dots, s_k, \lambda_{1:k+1}\right) \prod_{j=1}^{k+1} \pi(\lambda_j|\alpha, \beta) \times \pi(\beta|\xi, \delta) \\
&= \prod_{j=1}^{k+1} \left[\lambda_j^{(s_j - s_{j-1})} - \exp^{\lambda_j \sum_{i=s_{(j-1)+1}}^{s_j} y_i} \right] \\
&\times \prod_{j=1}^{k+1} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda_j^{\alpha-1} \exp\left(\beta\lambda_j\right) \times \frac{\delta^{\xi}}{\Gamma(\xi)} \beta^{\xi-1} \exp\left(\delta\beta\right).
\end{aligned}$$

The marginal distribution of $\pi(\beta|k, s_1, \dots, s_k, y_{1:d}, \lambda_{1:k+1}, \alpha, \xi, \delta)$ is

$$\begin{aligned}
\pi(\beta|k, s_1, \dots, s_k, y_{1:d}, \lambda_{1:k+1}, \alpha, \xi, \delta) &\propto \prod_{j=1}^{k+1} \left[\lambda_j^{(s_j - s_{j-1})} - \exp^{\lambda_j \sum_{i=s_{(j-1)+1}}^{s_j} y_i} \right] \\
&\times \prod_{j=1}^{k+1} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda_j^{\alpha-1} \exp\left(\beta\lambda_j\right) \times \frac{\delta^{\xi}}{\Gamma(\xi)} \beta^{\xi-1} \exp\left(-\delta\beta\right) \\
&\propto \prod_{j=1}^{k+1} \beta^{\alpha} \exp\left(\beta\lambda_j\right) \times \beta^{\xi-1} \exp\left(-\delta\beta\right) \\
&= \beta^{(k+1)\alpha+\xi-1} \exp\left(-\beta\left[\sum_{j=1}^{k+1} \lambda_j + \delta\right]\right).
\end{aligned}$$

This is the kernel of a gamma distribution with shape $(k+1)\alpha + \xi$ and rate $\sum_{j=1}^{k+1} \lambda_j + \delta$ and is updated once each iteration. The hazards are sampled from a gamma distribution $\lambda_j|\alpha, \beta, k \sim \mathcal{G}(\alpha + s_j - s_{j-1}, \beta^* + \sum_{i=s_{(j-1)+1}}^{s_j} y_i)$ with β^* the current value of β before the sampling of a new β.

One important practical point relates to the choice of ξ, δ when the data has a particular timescale. If the data is in years we should set β to have an expected value of 1 (assuming that in all cases α is set to 1) with variance 1 which is achieved by setting $\xi = 1, \delta = 1$ and follows immediately from the properties of the Gamma distribution. If the data is in days and we wish to retain the same prior we require β to have an expected value and variance of 365 which is achieved $\xi = 1, \delta = 1/365$.

B Piecewise Exponential Models Applied to HTA

B.1 Overview of Bagust and Beale Approach to Survival Extrapolation

The manuscript by Bagust and Beale (2014) primarily focuses on assessing cumulative hazard plots in order to identify a timepoint at which there is evidence of a long-term linear trend in the hazard (as previously described in our study). Related to this, they describe a situation where long-term trends in each arm of a clinical trial appear to exhibit very similar hazard rates. They suggest that, after the intervention treatment is completed, withdrawn, or ceases to deliver additional benefit, an identical long-term risk trajectory applies regardless of treatment, again assessed by cumulative hazard and post-progression survival plots for each arm.

Bagust and Beale also suggest that the hazard function can be conceptualized as arising from a mixture of populations each represented as having a different underlying hazard function. They reference previous work, which showed that a Weibull distribution with shape parameter less than 1 can be mathematically formulated as a mixture of exponential distributions (Jewell, 1982). Additionally, the exponential distribution itself arises as a mixture of Weibull distributions with fixed shape parameter less than 1 (implying a monotonically decreasing hazard). They suggest that identifying such subgroups may be useful "in lending credibility to some projective models, as well as in furnishing new hypotheses for targeting research to identify patients most likely to benefit from treatment".

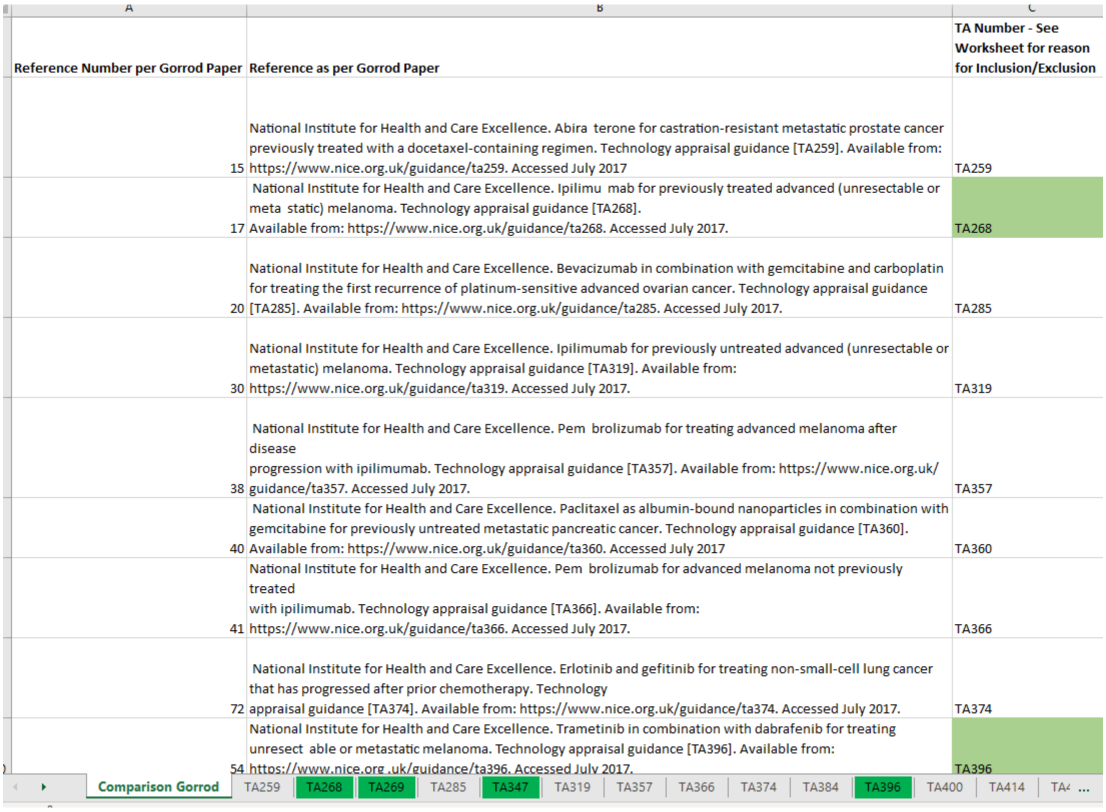
The manuscript also criticizes National Institute for Health and Care Excellence (NICE) Technical Support Document (TSD) 14 on a number of points (Latimer, 2013). In particular the authors suggest that the focus should be on identifying hypotheses which might lend credible extrapolations rather than assuming one of the "standard" parametric models will adequately predict the long-term survival. They also criticise the use of log cumulative hazard plots (rather than cumulative hazard plots), suggesting that the visual assessment of long-term constant hazards is confounded by the fact that both an exponential and Weibull survival model will produce straight lines when using log cumulative hazard plots. It should be noted that the author of NICE TSD 14 contested many of these criticisms in a response (Latimer, 2014).

B.2 Replication of Results Presented in Chapter 3

B.2.1 Details of Data Extraction from Technology Appraisals

Information used for the identification of the relevant Technology Appraisals (TAs) from the review by Gorrod et al. along with the relevant information extracted from them can be found in the excel file called `Summary of Piecewise TAs.xlsx` located within the `Files_Replicate_Analysis` folder Gorrod et al. (2019).

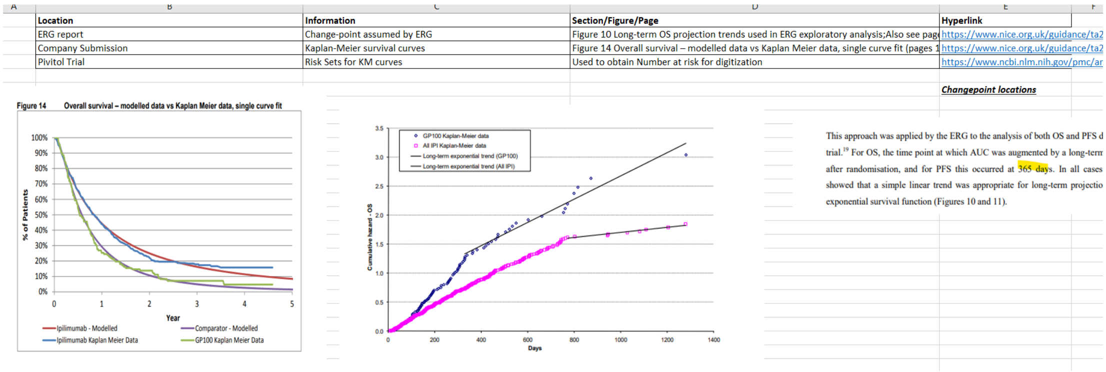
The first worksheet of `Summary of Piecewise TAs.xlsx` lists the TAs investigated (Figure B.1).



Reference Number per Gorrod Paper	Reference as per Gorrod Paper	TA Number - See Worksheet for reason for Inclusion/Exclusion
15	National Institute for Health and Care Excellence. Abira terone for castration-resistant metastatic prostate cancer previously treated with a docetaxel-containing regimen. Technology appraisal guidance [TA259]. Available from: https://www.nice.org.uk/guidance/ta259. Accessed July 2017	TA259
17	National Institute for Health and Care Excellence. Ipilimu mab for previously treated advanced (unresectable or meta static) melanoma. Technology appraisal guidance [TA268]. Available from: https://www.nice.org.uk/guidance/ta268. Accessed July 2017.	TA268
20	National Institute for Health and Care Excellence. Bevacizumab in combination with gemcitabine and carboplatin for treating the first recurrence of platinum-sensitive advanced ovarian cancer. Technology appraisal guidance [TA285]. Available from: https://www.nice.org.uk/guidance/ta285. Accessed July 2017.	TA285
30	National Institute for Health and Care Excellence. Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma. Technology appraisal guidance [TA319]. Available from: https://www.nice.org.uk/guidance/ta319. Accessed July 2017.	TA319
38	National Institute for Health and Care Excellence. Pem brolizumab for treating advanced melanoma after disease progression with ipilimumab. Technology appraisal guidance [TA357]. Available from: https://www.nice.org.uk/guidance/ta357. Accessed July 2017.	TA357
40	National Institute for Health and Care Excellence. Paclitaxel as albumin-bound nanoparticles in combination with gemcitabine for previously untreated metastatic pancreatic cancer. Technology appraisal guidance [TA360]. Available from: https://www.nice.org.uk/guidance/ta360. Accessed July 2017	TA360
41	National Institute for Health and Care Excellence. Pem brolizumab for advanced melanoma not previously treated with ipilimumab. Technology appraisal guidance [TA366]. Available from: https://www.nice.org.uk/guidance/ta366. Accessed July 2017.	TA366
72	National Institute for Health and Care Excellence. Erlotinib and gefitinib for treating non-small-cell lung cancer that has progressed after prior chemotherapy. Technology appraisal guidance [TA374]. Available from: https://www.nice.org.uk/guidance/ta374. Accessed July 2017.	TA374
54	National Institute for Health and Care Excellence. Trametinib in combination with dabrafenib for treating unresect able or metastatic melanoma. Technology appraisal guidance [TA396]. Available from: https://www.nice.org .uk/guidance/ta396. Accessed July 2017.	TA396

Figure B.1: Excel worksheet with Overview of all Technologies Appraisals considered

For each of the TAs listed in the first worksheet, a separate worksheet provides further information relating to whether or not the Bagust and Beale (B&B) approach was used along with the relevant location in the TA and an associated screengrab of the relevant information. In situations where the B&B approach was used, further information including the Kaplan-Meier survival functions (for generation of the pseudo-data) and location of the assumed change-point are highlighted. The locations (such as page numbers or section numbers) of all the extracted information within the TA are also recorded. Kaplan-Meier survival functions for any survival data made available after the original TA is presented along with a link to the relevant data-source. For an example of some of the data extracted from TA268 see Figure B.2.



Location	Information	Section/Figure/Page	Hyperlink
ERG report	Change-point assumed by ERG	Figure 10 Long-term OS projection trends used in ERG exploratory analysis;Also see pag	https://www.nice.org.uk/guidance/ta2
Company Submission	Kaplan-Meier survival curves	Figure 14 Overall survival – modelled data vs Kaplan Meier data, single curve fit (pages 1	https://www.nice.org.uk/guidance/ta2
Pivitol Trial	Risk Sets for KM curves	Used to obtain Number at risk for digitization	https://www.ncbi.nlm.nih.gov/pmc/ar

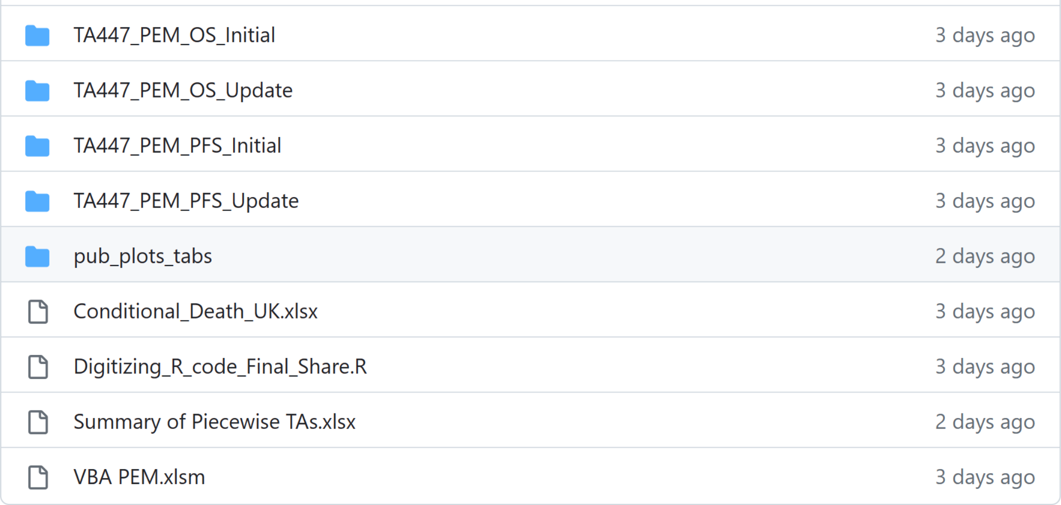
Changepoint locations

This approach was applied by the ERG to the analysis of both OS and PFS d
trial.[19] For OS, the time point at which AUC was augmented by a long-term
after randomisation, and for PFS this occurred at 365 days. In all cases
showed that a simple linear trend was appropriate for long-term projectio
exponential survival function (Figures 10 and 11).

Figure B.2: Excel worksheet with details extracted from TA268

B.2.2 Analysis of Extracted data and Simulation Studies using PiecewiseChangepoint package

All results from the manuscript can be replicated by locating the R script titled `Digitizing_R_code_Final_Share.R` within the `Files_Replicate_Analysis` folder. This script will produce relevant plots and tables in the folder named `pub-plots-tabs`, using the PiecewiseChangepoint package and associated R functions described in Section 8.1. A number of sub-folders are also contained within the `Files_Replicate_Analysis` folder and provide pseudo-patient data created from the Kaplan-Meier survival functions presented in the TAs (and publications providing later data cuts). These are named using the following structure: `TA_Treatment_Outcome_Datacut` and use a R function called `digitize` in addition to `survival.txt` and `nrisk.txt` files in these folders to create the associated dataset. Survival models are fit to these datasets to generate the results in the main manuscript and figures in Section B.3.



TA447_PEM_OS_Initial	3 days ago
TA447_PEM_OS_Update	3 days ago
TA447_PEM_PFS_Initial	3 days ago
TA447_PEM_PFS_Update	3 days ago
pub_plots_tabs	2 days ago
Conditional_Death_UK.xlsx	3 days ago
Digitizing_R_code_Final_Share.R	3 days ago
Summary of Piecewise TAs.xlsx	2 days ago
VBA PEM.xlsm	3 days ago

Figure B.3: Overview of File Structure

Separately, a file called `Simulation Study.R` provides the code required to produce the results presented in Section 3.6.

B.3 Supplementary analysis of updated data for TA396, TA428 and TA447

In Figures B.4 and B.5 we compare the original data from TA396 to an updated pooled analysis of the COMBI-v and COMBI-d data for both OS and PFS outcomes respectively. In all figures in this section the Kaplan-Meier survival function before dashed vertical line indicates earlier data-cut, while Kaplan-Meier survival function afterwards is the long-term follow up. For Figure B.5 due to the large number of change-points only the final change-point is presented.

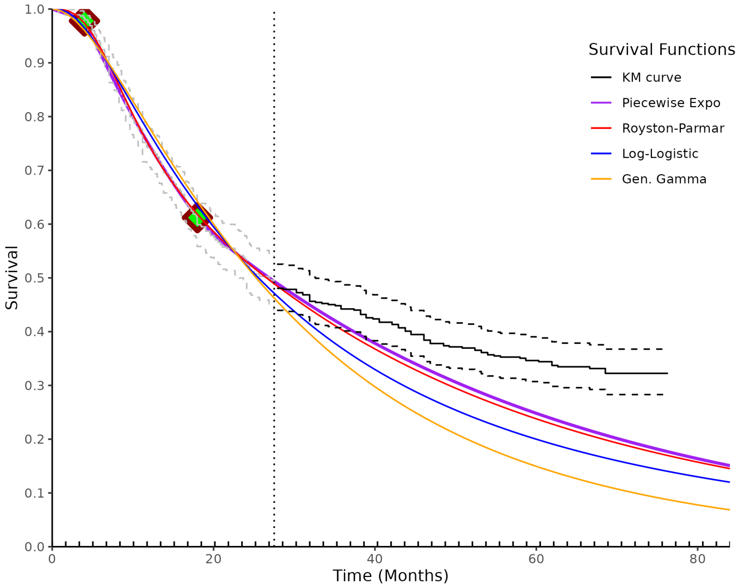


Figure B.4: Long-term survival probabilities (OS) for various models compared to long-term data from the COMBI-v and COMBI-d trials.

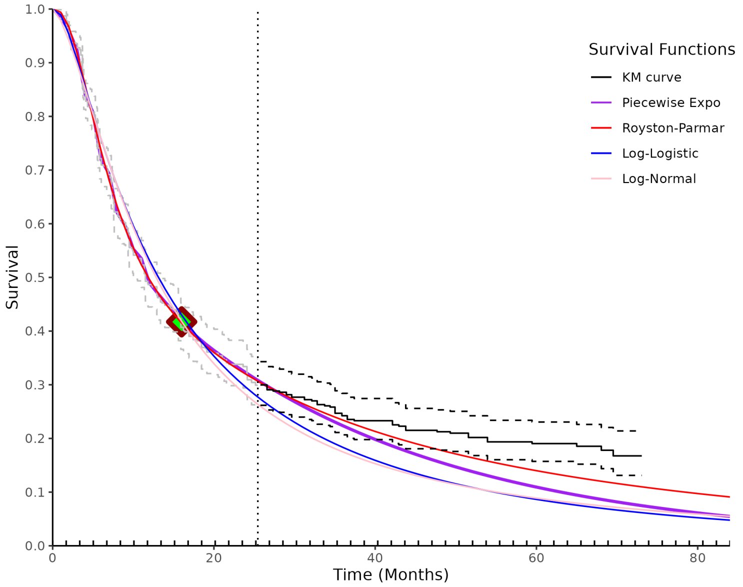


Figure B.5: Long-term survival probabilities (PFS) for various models compared to long-term data from the COMBI-v and COMBI-d trials.

In Figure B.6 we compare the original data from TA447 to that in TA531, which superseded it. It should be noted that most probable change-point model was the no-change-point i.e. exponential model.

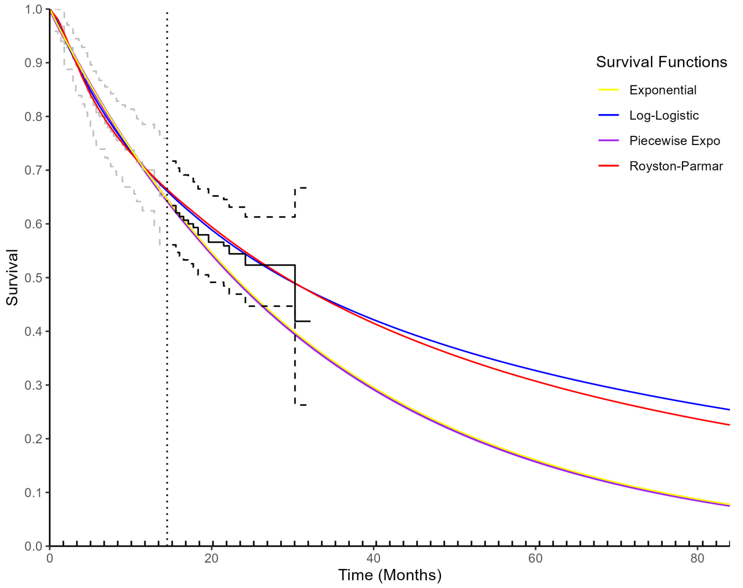


Figure B.6: Long-term survival probabilities (OS) for various models compared to long-term data from TA531 (Update to TA447 data). Kaplan-Meier survival function before dashed vertical line indicates earlier data-cut, while Kaplan-Meier survival function afterwards is the long-term follow up.

C Psuedo Code for Weibull Change-point Model

The pseudo-code presented below shows how a Weibull change-point model could be implemented in the JAGS statistical software. Modifications to include covariates are relatively straightfoward.

```
model {
  # Assuming a one change-point Weibull model with no covariates
  # Uniform prior for shape and scale.
  # Constant for zerors trick
  C <- 10000
  N_CP <- 1
  cp[1] = 0  # Should be zero
  cp[2] ~ dunif(0, max(time)) # Alternatives possible
  cp[3] = 100  # mcp helper value. very large number
  #Prior for the model parameters
  for(k in 1:(N_CP+1)){
      shape[k] ~ dunif(0,10)
      scale[k] ~ dunif(0,10)
  }
  # Model and likelihood
  for (i in 1:N) {
    for(k in 1:(N_CP+1)){
    #variable which gives the difference between the two intervals if
    ↪  time[i]>cp[k+1]
    #(i.e. cp[k+1] -cp[k]) or time between time[i] and cp[k]
    X[i,k] = max(min(time[i], cp[k+1]) - cp[k],0)
    #Indicator variable which highlights which interval time is in
    X_ind[i,k] = step(time[i]-cp[k])*step(cp[k+1]-time[i])

    log_haz_seg[i,k] <-
    ↪  log(shape[k]*scale[k]*pow(time[i],shape[i,k]-1))*X_ind[i,k]
    cum_haz_seg[i,k] <- scale[i,k]*pow(X[i,k]+cp[k],shape[i,k]) -
                        scale[i,k]*pow(cp[k],shape[i,k])
    }
    log_haz[i] <- sum(log_haz_seg[i,])
    cum_haz[i] <- sum(cum_haz_seg[i,])
    loglik[i] = status[i]*log_haz[i] - cum_haz[i]
    #Zero Trick
    zero[i] ~ dpois(C - loglik[i])
  }
}
```

Listing 8: Pseudo-Code for JAGS Change-point Model

D Expert Opinion in Survival analysis

D.1 Validation of approach

We compare the results of Singpurewalla and Song (1988) to our proposed method. Singpurewalla and Song (1988) consider a Weibull distribution with a proportional hazards (PH) parameterization. The median survival time is $t_{0.5} = \frac{\log(2)}{m}^{\frac{1}{a}} = \kappa$. Re-expressing the distribution in terms of κ, we obtain survival function $S(t) = \exp\Big\{ - \log(2)\big(\frac{t}{\kappa}^{a}\big)\Big\}$ and hazard function $h(t) = \frac{\log(2)at^{a-1}}{\kappa^a}$. From this we can obtain the likelihood of this data using the expressions in Section 7.2.1.

The expert belief about κ is characterized by the location or mean l and standard deviation s. Singpurewalla and Song (1988) also consider additional parameters c and v which can be used to calibrate the expert's opinion about l and s, however, in the case the analyst does not wish to modulate the expert's opinion then $c = 1$ and $v = \frac{1}{2}$. By invoking some mild assumptions Singpurewalla and Song (1988) state that

$$[c^2v/(sl^2)]\kappa^2 \approx \chi^2((v/s)+1).$$

Using the change of variables technique the density for the median survival is

$$p(\kappa|l,s,c,v) = \chi^2(\kappa^2(c^2v)/(sl^2)|(v/s)+1) \times 2\kappa(c^2v)/(sl^2).$$

The scale parameter is assumed to have a gamma prior $a \sim \mathcal{G}(\alpha,\beta)$, with the parameters of this distribution specified by the expert. Singpurewalla and Song (1988) describe a Bayes estimator for the parameters using some approximations, however, it is straightforward to use JAGS or Stan to obtain the complete posterior distribution. Using simulated data they provide in their paper, they set $l = 500$, $s = 200$, $\alpha = 6.25$ and $\beta = 12.5$ and did not assume any modulation of the expert's opinion. This gives a prior for the median survival (termed Original Prior) in Figure D.1 and posterior survival functions in Figure D.2a. The fact that the posterior distributions are very similar to the analysis without any expert opinion is unsurprising as the original prior had a significant probability within the 95% confidence interval implied by the data alone [8909 - 22188]. In a second example we adjust the prior belief of the expert to yield a much lower median value (termed adjusted prior in Figure D.1) and see that the mean survival posterior for both approaches incorporating expert opinion are very similar and as expected, outside the confidence interval for the median (Figure D.2b). This highlights that our proposed approach is consistent with previous methods.

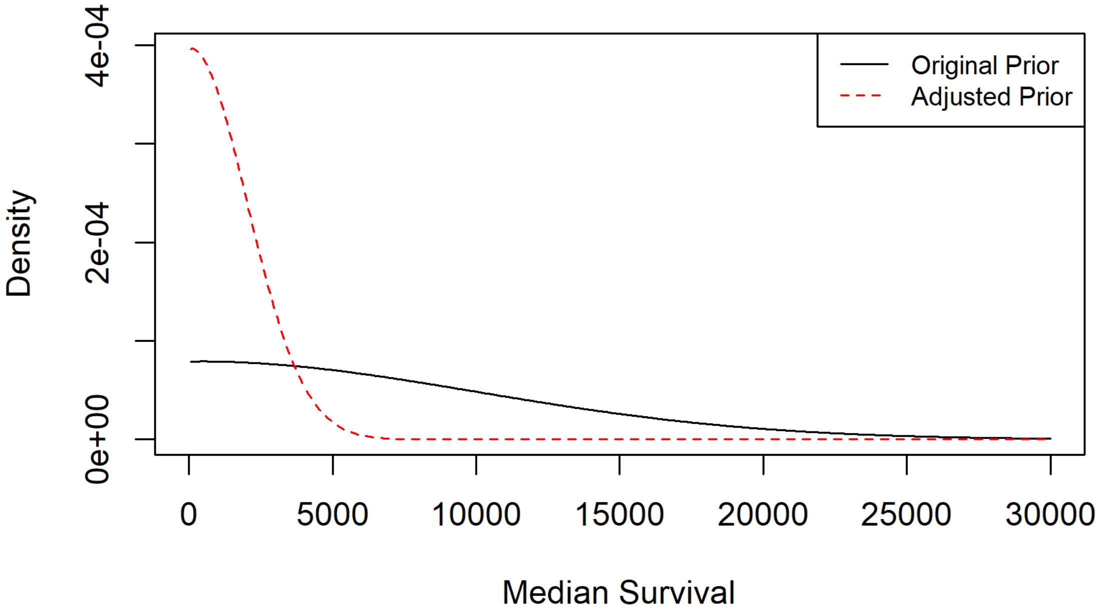


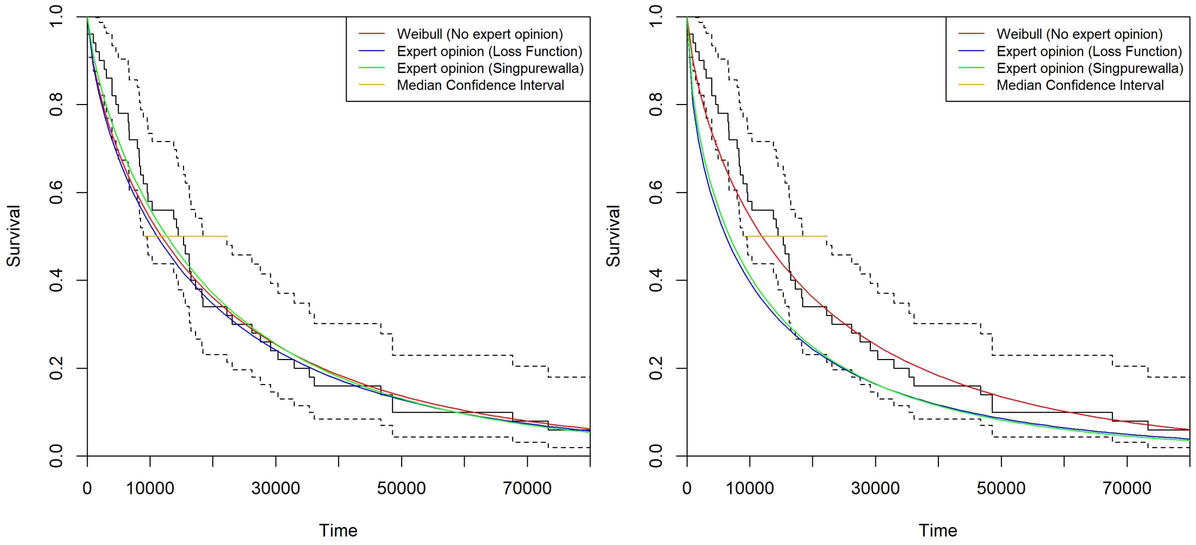
Figure D.1: Original prior used by Singpurewalla and Song (1988) and adjusted prior

As can be seen from this example, specifying expert opinion in which the elicited belief is a function of a chi-squared distribution is not particularly intuitive, however, it highlights that the proposed approach is consistent with previous methods and can be easily adjusted to include prior beliefs (using any distribution) on any quantity of interest.

D.2 Technical Details for models fit in JAGS

We could not fit the Gamma, Gompertz and generalized Gamma models in Stan and we instead fit these models in JAGS. We describe how we analytically evaluate the expected survival for the Gompertz and generalized Gamma distributions.

The expectation of Gompertz distributed random variable with shape parameter $a^* = 1/b$ and rate parameter $b^* = a/b$ is $E[T] = b \exp aE_a(-1)$ where $E_a(t) = \int_1^\infty u^t \exp(-au)du$



(a) Survival functions under the original prior (b) Survival functions under the adjusted prior

Figure D.2: Comparison of Approaching incorporating expert opinion

(Siegrist, 2023). We note that $E_a(t) = \Gamma(0, a)$, the upper incomplete gamma function (Weisstein, 2021). We need to approximate this function, and note that by definition the gamma function ($\Gamma(x)$) is the sum of the upper ($\Gamma(x, a)$) and lower ($\gamma(x, a)$) incomplete gamma functions $\Gamma(x) = \Gamma(x, a) + \gamma(x, a)$. Hence, $\Gamma(0, a) = \lim_{x \to 0}(\Gamma(x) - \gamma(x, a))$ where for practical purposes we set x = 0.0001. By definition we can compute the lower incomplete gamma function as a product of the gamma function and the cumulative distribution function of the gamma distribution with the following parameters: $\Gamma(x)F(x = a; \alpha = s, \beta = 1)$. We finally have the expected survival as $(1/a^*) \exp\{b^*/a^*\}\Gamma(0, b^*/a^*)$.

For Generalized Gamma the parameterization in JAGS is slightly different to Stacy (1962), with $b \times r = d$, $b = p$ and $\lambda = 1/a$, which gives the the mean as $\frac{\Gamma((b \times r+1)/b)}{\lambda\gamma(r)}$. For consistency of results with the `flexsurv` package we have $\mu = -\log(\lambda) + \log(r)/b$, $\sigma = 1/(b * \sqrt{r})$ and $Q = \sqrt{1/r}$.

D.3 Effective sample size for Pareto distribution

Another survival model for which we can use the prior distribution to encode information about the effective sample size. Suppose we have a Pareto distribution with a fixed scale θ ($\theta < t$) but shape λ is Gamma distributed. Using the substitution $y = \ln(t/\theta)$, it can be shown that y, unconditional on λ, is Lomax distributed with parameters α, β (McNulty, 2021), with α being the effective sample size n_e. Letting t^* be the timepoint at which the expert's opinion is elicited.

$$\begin{aligned}
\theta &\sim \mathcal{U}(0, t^*) \\
\beta &= \frac{\log(t^*/\theta)}{(\hat{S(t^*)}^{-1/n_e} - 1)} \\
\lambda &\sim \mathcal{G}(n_e, \beta) \\
T &\sim P(\lambda, \beta)
\end{aligned}$$

D.4 Simulation study - Effect of priors on posterior survival when including expert opinion

As noted in the main text we wished to assess if the weakly informative priors typically used in Bayesian analysis could conflict with the information provided by the expert. To investigate this, we conducted a simulation study comparing the posterior survival of the models with expert opinion under two specifications of weakly informative priors; one in which the priors for all parameters were uniform and alternatively where the priors for parameters had normal or gamma distributions (the latter for parameters which are constrained to be positive). In addition to the priors having a different parametric form, a further difference was that the standard deviations for normal and gamma distributions were relatively low for weakly informative priors. For example, the log of the scale parameter for the Weibull model was a normal distribution with mean 0 and standard deviation 1 and the shape parameter was a gamma distribution with both parameters equal to 1. This is in contrast to the standard deviations of the uniform priors, which were typically >28. In the simulation study we generated data from a Weibull (proportional hazards) model for a variety of parameters and sample sizes (n_{samp} = 30,50,100) with a maximum follow-up of 2 years. We incorporated different values of expert opinion in terms of mean and standard deviation at multiple timepoints, assuming that the expert's opinion was a normal distribution. Taking all combinations of the parameters described in Table D.1 produced 324 simulations across each of the parametric models.

To assess the similarity of the survival functions under each prior specification we evaluated the posterior median restricted mean survival time (RMST) until a timepoint of 15 years for each model. To provide a measure of similarity on a comparable scale, we estimated the ratio $\frac{\min(RMST_{\text{uniform}}, RMST_{\text{non-uniform}})}{\max(RMST_{\text{uniform}}, RMST_{\text{non-uniform}})}$, with $\text{RMST}_{\text{uniform}}$ and $\text{RMST}_{\text{non-uniform}}$ denoting the RMST under each prior specification. Overall, all models had very high RMST ratio values with median values of 0.99 even at a sample size of 30 as shown in Table 3. Results for larger sample sizes were even larger.

A simulation study using the same specifications as described in Table D.2 compared the

Table D.1: Parameters used in simulation study to investigate the effect on model priors when including expert opinion

Parameter	Values
Shape	0.75, 1, 1.25
Scale	0.25,0.5,0.75, 1
Sample Size	30,50,100
Mean value of $S(t_1^*)$	0.1, 0.3
Mean value of $S(t_2^*)$	0.05, 0.1
SD of expert's opinion	0.025, 0.05, 0.1

SD of expert's opinion† 0.025, 0.05, 0.1;
$S(t_1^*)S(t_2^*)$ denotes the survival at 4 and 10 years – only evaluated scenarios in which $S(t_1^*) > S(t_2^*)$
† Standard deviation (SD) of the expert opinion was equal across both timepoints

Table D.2: RSMT ratios for each parametric model with dataset of 30 observations in both simulation studies

Model	Median RMST Ratio – Uniform vs Normal/Gamma Priors	Median RMST Ratio – Uniform vs Penalized Maximum Likelihood Estimates
Exponential	0.99	0.99
Weibull	0.99	0.97
log-Logistic	0.99	1
log-Normal	0.99	0.99
Royston-Parmar	1	0.98
Gompertz	0.99	0.97
Gamma	0.99	0.97
Gen. Gamma	0.99	0.98

Bayesian approach (with uniform priors) to the estimates derived from the penalized maximum likelihood approach (which does not require the specification of a prior). If the distribution representing the expert opinion was not multi-modal, the ratios of RMST were very close to 1 even at sample sizes of 30. Based on these results it can be concluded that weakly informative prior distributions do not conflict with the information provided by the expert.

For illustration, Figure D.3 below shows examples of the posterior survival functions (along with 95% intervals as dashed lines) for the Gompertz, Weibull, Royston-Parmar spline model and generalized gamma for scenarios in which the sample size was 30 observations. The RMST ratio is provided in the title of each plot. Expert opinion is indicated by the dashed vertical lines at times 4 and 10 and includes situations where the mean values of $S(t^*)$ are both high and low and also informative and vague (small and large standard deviations).

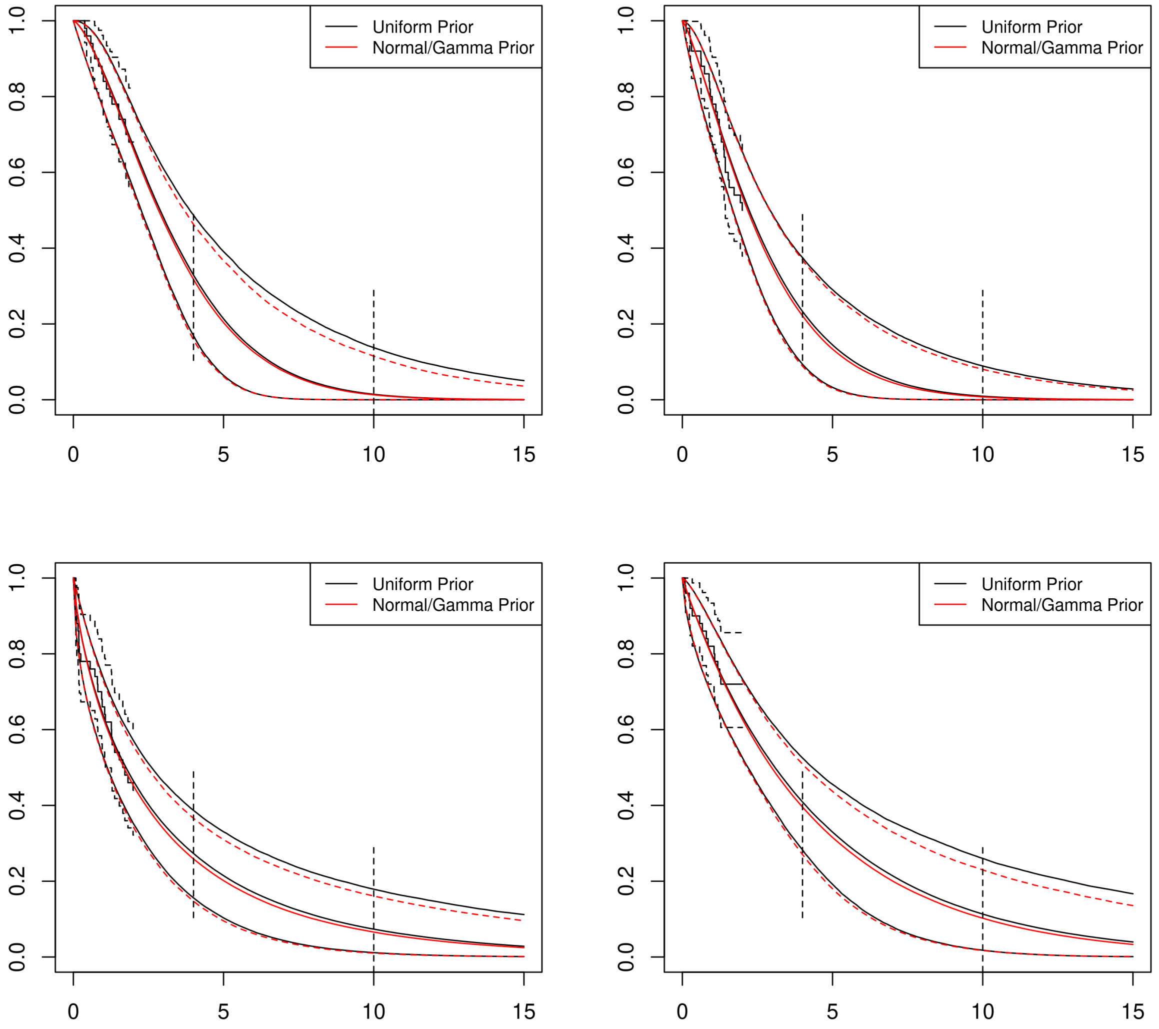


Figure D.3: Survival functions when under different specifications of minimally informative priors. Shown clockwise are Gompertz, Log-Normal, Royston Parmar Spline and Weibull models fit to different simulated datasets.

Similar illustrations are provided for the comparison between the Bayesian models with expert opinion and the penalized maximum likelihood approach for gamma, log-normal, Royston-Parmar and Weibull models (Figure D.4). Also included for reference are the Bayesian and maximum likelihood models without any expert opinion.

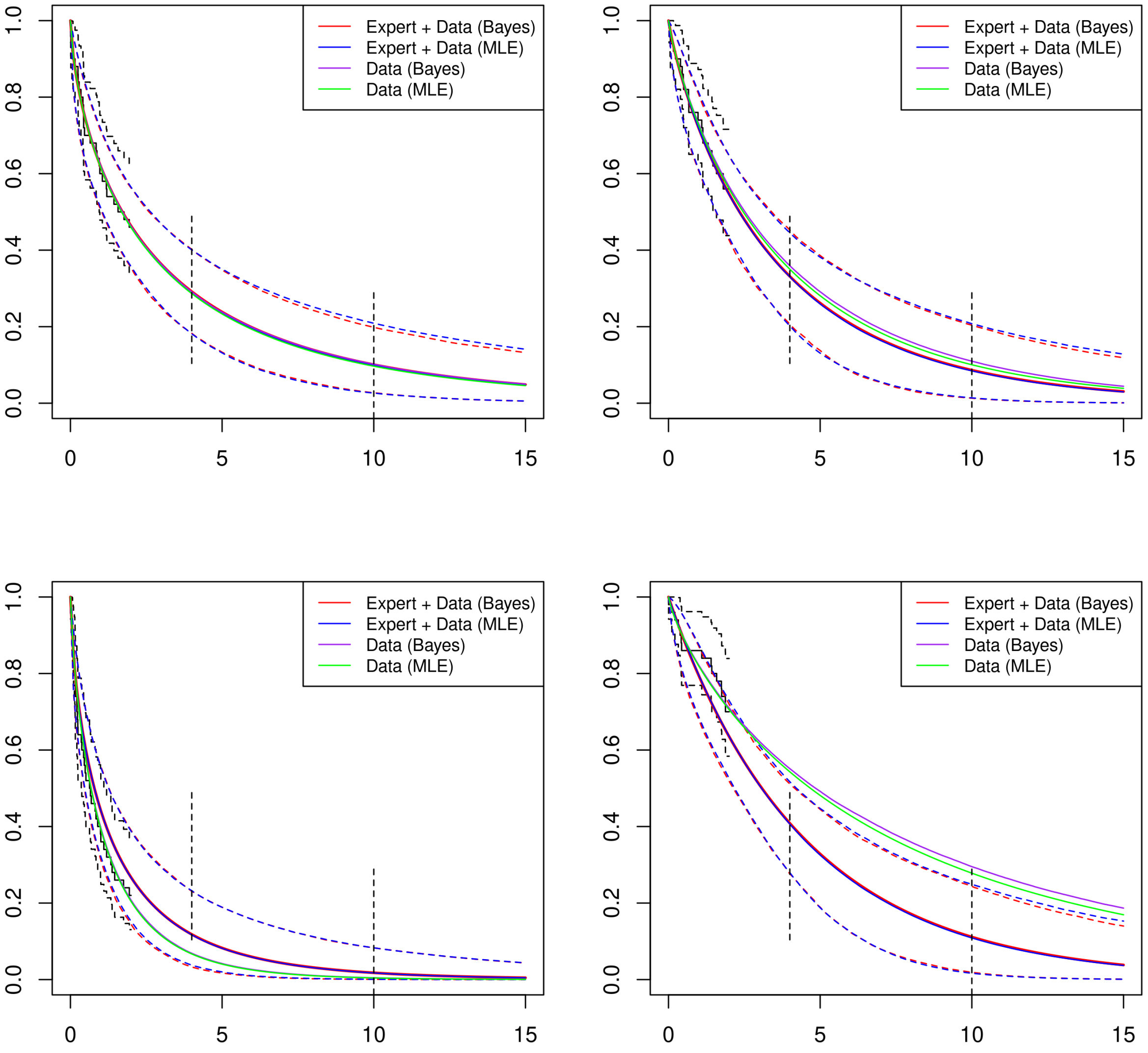


Figure D.4: Survival functions comparing Bayesian and Penalized likelihood approaches (scenarios with and without opinions). Shown clockwise are Gompertz, Log-Normal, Royston Parmar Spline and Weibull models fit to different simulated datasets.

D.5 Simulation study – Impact of Bias in Expert Opinion on Extrapolated Survival

As noted in main text, it is almost certain that the opinions elicited from the expert (and parameterized as probability distributions) will not be centred on the true value. Considering beliefs elicited about survival probabilities at timepoints, if the expected value of an expert's opinion is different from the true survival at a particular timepoint, then the expert's opinion is biased relative to the true survival. However, in many situations it can still be closer (on average) to the true survival function than using the data alone. To make this statement more concrete, consider an example in which 30 observations are generated by a Weibull probability distribution with shape equals 1, scale equals 0.1

(proportional hazards parameterization) and a maximum follow up time of 2 years. If the expert assumes that their belief about survival at 10 years is characterized by a normal distribution with a mean of 0.46 and standard deviation of 0.05, the expected value of their opinion is 25% above the true value of 0.367, i.e. biased by a factor of 1.25. If we fit a Weibull model to the data, the maximum likelihood estimate will (on average) provide an approximately unbiased survival function, however, because we have a limited sample size and do not observe data after 2 years the estimate is associated with a significant degree of uncertainty. In contrast, the survival function (at the restricted maximum likelihood estimate) obtained from including the expert's opinion will be biased but have a considerable reduction in uncertainty and will be on average closer to the true survival function. To produce stable results, 500 datasets are simulated under the conditions described above and models fit by (restricted) maximum likelihood including and excluding the expert's opinion. Figure D.5 presents the median (solid line) and 95% quantiles (dashed lines) of the survival estimated with (purple) and without (red) expert opinion. To be clear, these are the quantiles of the survival over the 500 datasets at the restricted and regular maximum likelihood estimates and do not refer to the confidence intervals from a given model.

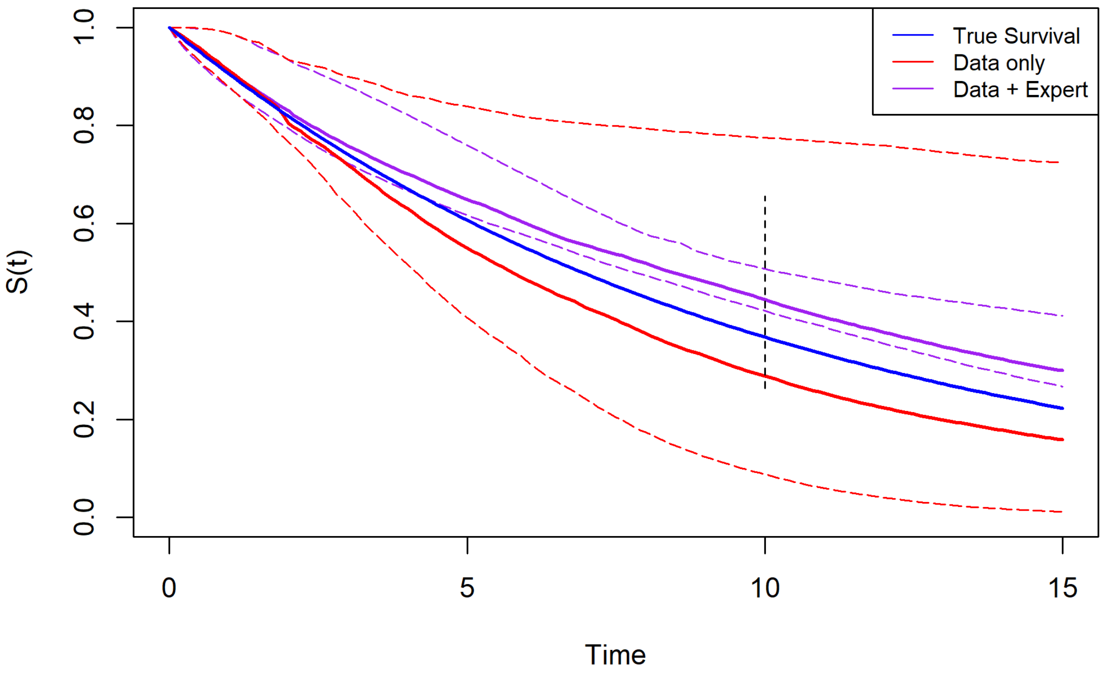


Figure D.5: Estimation of expected survival functions with and without expert opinion

From Figure D.5 we get a sense of the bias-variance trade off. Although the survival at 10 years estimated by the penalized maximum likelihood estimators are always above the true survival for each of the 500 datasets (ranging from 0.42-0.50), they are less than 0.13 from the true value in 97.5% of the datasets. In contrast when using the data alone the 95% interval for the survival is 0.08-0.78 meaning that in 5% of the datasets either underestimated the survival by ≥ 0.286 or overestimated it by ≥ 0.40. In order to get a numerical estimate of the bias-variance trade off, we estimated the mean squared error (MSE) for the restricted mean survival time (RMST), up to 15 years: $MSE = (RMST - \widehat{RMST})^2$ with RMST referring to the true value and $\widehat{RMST}$ the estimate based on the models, for both the model with and without expert opinion (i.e. data alone) in each of the 500 simulations. To get a single number to compare the results between both models, we evaluated the mean of the difference (across the 500 simulations) between the MSE from the model without expert opinion and MSE with expert opinion. Values >0 mean that on average the MSE for the model with expert opinion was lower than that without expert opinion. In our example the median MSE was 5 without using expert opinion, and 0.94 using expert opinion. Across the 500 simulations the average difference of MSE without expert opinion against MSE with expert opinion was 5.55. We also evaluated the absolute difference of the RMST, $|RMST - \widehat{RMST}|$ for both the models with and without expert opinion and evaluated the mean difference. Using this measure (the absolute deviation rather than squared deviation) places a lower penalty than MSE on having values further away from the true value. In this situation the expected difference in absolute difference of RMST was 1.22, considerably lower than the equivalent value based on the squared deviation but nevertheless substantial, as the true RMST was 7.77. We expect a variety of parameters to influence the MSE with and without expert opinion, in particular degree of bias, strength of belief (as indicated by the standard deviation) and the sample size. These factors and others relating to the follow-up time, parameters for the Weibull distribution and timepoints of expert's opinion (parameterized as a normal distribution) are presented in Table D.3. We considered specifications based on each of the combinations of parameters, to yield 576 specifications.

Table D.3: Parameters for simulation study comparing predictive accuracy of survival models including expert opinion with models estimated based solely on the observed data.

Parameter	Values
Shape	0.75, 1, 1.25
Scale	0.25,0.5,0.75, 1
Sample Size	30,50,100
Standard Deviation of expert's opinion*	0.1,0.05,0.025
Timepoints for expert opinion	4 only, 10 only, 4 and 10
Longest follow up time (i.e. observations after this assumed censored)*	2, 4
Bias factor of expert – factor by which the expert under/overestimates true survival	0.6,.75, 1.25

* If follow up time was 4 years, then expert opinion was only incorporated at 10 years

Overall, only 70 of the 576 (12%) simulations had an expected difference in MSE < 0, denoting that on average the MSE with expert opinion was worse than without expert opinion. In the case of absolute deviation this number increased to 118 simulations (20%). Considering the results summarized by bias factor of expert and standard deviation of expert opinion (Table D.4), the percentage of scenarios in which (expected) absolute deviation from RMST was lower with expert opinion was above 70% for all scenarios in which the bias factor of the expert was between 0.75-1.25 and suggests that expert opinions within this range improve the prediction of long-term survival outcomes.

Perhaps unsurprisingly the situations where the expert opinion had a higher (worse) MSE included when the expert was biased and the true survival was quite high i.e. above 40% for years 4 or 10. This is because we included a relative bias, assuming the survival is 1.25 times the true survival at 50% results in a greater absolute error than when it is only 25%. From the perspective of including expert opinion, the worst results were obtained for bias factors farther away from 1, large sample sizes, and in which the confidence of the expert was high e.g. standard deviation equal to 0.025. Lower MSE was achieved in situations where expert opinion was incorporated at the 10 year timepoint rather than the 4 year timepoint.

Table D.4: Percentage of scenarios in which model with expert opinion performed better than model based on data alone (Models and data assumed to be from a Weibull distribution)

Bias factor of expert	SD of expert's opinion	% Scenarios in which (average) squared deviation from RMST was lower with expert opinion	% Scenarios in which (average) absolute deviation from RMST was lower with expert opinion	Number of Scenarios
0.6	0.025	60%	40%	48
0.6	0.05	67%	50%	48
0.6	0.1	85%	65%	48
0.75	0.025	85%	75%	48
0.75	0.05	90%	81%	48
0.75	0.1	96%	94%	48
0.9	0.025	100%	100%	48
0.9	0.05	100%	100%	48
0.9	0.1	100%	100%	48
1.25	0.025	83%	73%	48
1.25	0.05	88%	77%	48
1.25	0.1	100%	100%	48

SD - Standard Deviation

It is also worth highlighting that results of the simulation study presented in Table D.4 was based on the assumption that the true model, a Weibull distribution, was selected. We repeated the simulation study assuming that the data were still generated by a Weibull distribution but that a log-normal distribution, i.e., an incorrect parametric model was fit to the data instead. Table D.5 shows that for all combinations of bias and standard deviation of expert's opinion, the (expected) absolute deviation from RMST was lower with expert opinion in more than 70% of scenarios, highlighting that the inclusion of expert opinion can make extrapolation of survival outcomes more robust to misspecification of the parametric model.

Table D.5: Percentage of scenarios in which log-normal model with expert opinion performed better than log-normal model based on data alone assuming data was generated by a Weibull distribution

Bias factor of expert	SD of expert's opinion	% Scenarios in which (average) squared deviation from RMST was lower with expert opinion	% Scenarios in which (average) absolute deviation from RMST was lower with expert opinion	Number of Scenarios
0.6	0.025	81%	77%	48
0.6	0.05	94%	83%	48
0.6	0.1	100%	100%	48
0.75	0.025	100%	100%	48
0.75	0.05	100%	100%	48
0.75	0.1	100%	100%	48
0.9	0.025	100%	100%	48
0.9	0.05	100%	100%	48
0.9	0.1	100%	100%	48
1.25	0.025	81%	81%	48
1.25	0.05	81%	81%	48
1.25	0.1	83%	81%	48

SD - Standard Deviation

D.6 Frequently used Parametric Survival Models in HTA

Table D.6 presents the parameterizations of commonly used survival models. The parameter of primary interest (as per the `flexsurv` package) is colored in red, known as the location parameter and typically governs the mean or location for each distribution. The other parameters are ancillary parameters that determine the shape, variance, or higher moments of the distribution. These parameters impact the hazard function, which can take a variety of shapes depending on the distribution.

In many situations covariates (i.e. treatment indicator) are included on the location parameters and result in proportional hazard (PH) or accelerated time factor (ATF) models. Less frequently covariates can be included on the ancillary parameters which allow for more flexible modelling of the data, however, the models will no longer have the PH/ATF interpretations.

Table D.6: Standard Parameterization of Parametric Survival Models

	PDF	CDF	Hazard	Parameters
Exponential	$\lambda e^{-\lambda t}$	$1 - e^{-\lambda t}$	λ	$rate = \lambda > 0$
Weibull (AFT)	$\frac{a}{b}\left(\frac{t}{b}\right)^{a-1} e^{-(t/b)^a}$	$1 - e^{-(t/b)^a}$	$1\ \frac{a}{b}\left(\frac{t}{b}\right)^{a-1}$	shape $= a > 0$ $scale = b > 0$
Weibull (PH)[2]	$amt^{a-1}e^{-mt^a}$	$1 - e^{-mt^a}$	amt^{a-1}	shape $= a > 0$ $scale = m > 0$
Gompertz	$be^{at} \exp\left[-\frac{b}{a}(e^{at} - 1)\right]$	$1 - \exp\left[-\frac{b}{a}(e^{at} - 1)\right]$	be^{at}	shape $= a \in (-\infty, \infty)$ $rate = b > 0$
Gamma[3]	$\frac{b^a}{\Gamma(a)} t^{a-1}e^{-bt}$	$\frac{\gamma(a,bt)}{\Gamma(a)}$	f(t)/S(t)	shape $= a > 0$ $rate = b > 0$
Lognormal	$\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}$	$\Phi\left(\frac{\ln t - \mu}{\sigma}\right)$	f(t)/S(t)	$meanlog = \mu \in (-\infty, \infty)$ sdlog $= \sigma > 0$
LogLogistic	$\frac{(a/b)(t/b)^{a-1}}{(1+(t/b)^a)^2}$	$\frac{1}{(1+(t/b)^a)}$	$1 - \frac{(a/b)(t/b)^{a-1}}{(1+(t/b)^a)}$	shape $= a > 0$ $scale = b > 0$
Generalized Gamma[3,4]	$\frac{\lvert Q\rvert(Q^{-2})^{Q^{-2}}}{\sigma t \Gamma(Q^{-2})} \exp\left[Q^{-2}\left(Qw - e^{Qw}\right)\right]$	$\frac{\gamma(Q^{-2},u)}{\Gamma(Q^{-2})}$ if $Q \neq 0$ $\Phi(w)$ if $Q = 0$	f(t)/S(t)	$mu = \mu \in (-\infty, \infty)$ sigma $= \sigma > 0$ Q $= Q \in (-\infty, \infty)$
Royston-Parmar Splines				See Royston and Parmar (2002)

[1] Red colour refers to location parameter.

[2] The proportional hazard (PH) model is a reparameterization of the accelerated failure time (AFT) model with $m = b^{-a}$.

[3] $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is the gamma function, $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ is the lower incomplete gamma function.

[4] $w = (log(t) - \mu)/\sigma$, $u = Q^{-2}e^{Qw}$ and Φ is the cumulative normal distribution function.

[5] $w = (log(t) - \mu)/\sigma, \delta = (q^2 + 2p)^{1/2}$, $m_1 = 2(q^2 + 2p + q\delta)^{-1}$, $m_2 = 2(q^2 + 2p - q\delta)^{-1}$ and $B()$ is the beta function.