The use of real-world time-to-event data to inform Health Technology Appraisal: applications in cost-effectiveness modelling



Ian Koblbauer

Harris Manchester College

Doctor of Philosophy

Clinical Epidemiology and Medical Statistics

Trinity Term 2024

Abstract

The use of real-world time-to-event data to inform Health Technology Appraisal: applications in cost-effectiveness modelling

Ian Koblbauer, Harris Manchester College, Trinity Term 2024

Cost-effectiveness analysis is used in Health Technology Appraisal (HTA) to inform reimbursement decisions. With growing interest in utilising real-world data (RWD) for HTA, research investigating the impact of approaches for integrating RWD in cost-effectiveness analysis is needed. In my thesis, I investigated approaches for integrating real-world time-to-event (TTE) data in HTA to supplement trial-based parameters.

First, I conducted a literature review to identify uses of RWD to generalise cost-effectiveness analyses which used inputs from clinical trials. I found that baseline risk was the most frequently occurring parameter for integrating RWD, however systematic or consistent approaches for integrating real-world baseline survival were lacking. Next, I proposed approaches for integrating aggregate real-world baseline TTE data with trial-derived relative effects when faced with complex hazards. I evaluated impacts of various approaches on cost-effectiveness estimates and found that accurately capturing baseline survival was a substantial driver of cost-effectiveness estimates. Thereafter, I investigated the use of individual-level RWD, where data could not be fully adjusted to represent a target clinical trial population, to inform survival modelling approaches in HTA. I explored potential impacts on decision-making using a motivating example and found that the use of RWD provided valuable information to inform selection of preferred survival modelling approaches, which substantially impacted cost-effectiveness estimates. Next, I further explored the use of RWD to inform survival model selection, where censoring distributions, sample sizes, and follow-up times typically observed in clinical trials were introduced. I found that standard parametric models, rather than spline-based models, most often performed best under these circumstances.

My findings support the use of real-world TTE data as a valuable supplement to clinical trial data to inform decision-making, provided that approaches are justified, and limitations are acknowledged. Further research into approaches for integrating real-world TTE data in cost-effectiveness modelling for HTA is warranted.

To Bethany

Acknowledgements

First and foremost, I would like to thank my three supervisors, Rafael Pinedo-Villanueva, Dani Prieto-Alhambra, and Ed Burn, whose time, support, and combined knowledge made this thesis possible. I thank Danielle Newby and the other EHDEN team members for the many enlightening discussions surrounding the use of real-world data in HTA. I thank Xihang Chen and Hez Omulo for their guidance with accessing databases, and Spyros Kolovos for the many helpful discussions during the early stages of the thesis. I also thank my internal examiners at the University of Oxford, Jonathan Cook and Ramon Luengo-Fernandez, and my external examiner, Manuel Gomes, for their time taken to assess my work. I thank Afsie Sabokbar and Sam Burnell for all their student support throughout my studies. I also thank the people at the Clarendon Fund for their funding support and the many members of the University of Oxford, NDORMS, and HMC that make day-to-day operations and studying at the University of Oxford possible.

Lastly, I would like to thank my family, friends, and colleagues, past and present, for all their support. Most importantly, I am eternally grateful to my wife, Naomi, for all her patience, sacrifices, and support, as well as my young daughter, Bethany, who never fails to put a smile on my face.

Abbreviations

AIC: Akaike information criteria

AUC: area-under-the-curve

BIC: Bayesian information criteria

CPRD: Clinical Practice Research Datalink

DSU: Decision Support Unit

EAG: evidence assessment group

ECOG: Eastern Cooperative Oncology Group

EHDEN: European Health Data and Evidence Network

GP: general practitioner

HTA: Health Technology Appraisal

ICER: incremental cost-effectiveness ratio

KM: Kaplan-Meier

LTFU: long-term follow-up

LYG: life-years gained

NICE: National Institute for Health and Care Excellence

OMOP CDM: Observational Medical Outcomes Partnership Common Data Model

PFS: progression-free survival

PH: proportional hazards

QALY: quality-adjusted life-years

RCS: restricted cubic spline

RCT: randomised controlled trial

rMSE: root mean squared error

RMST: restricted mean survival time

RWD: real-world data

RWE: real-world evidence

SEER: surveillance, epidemiology, and end results program

TSD: technical support document

TVC: time-varying covariate

UK: United Kingdom

US: United States of America

Contents

1 Introduction 1
1.1 The use of cost-effectiveness analysis for decision-making 1
1.2 The use of real-world data in Health Technology Appraisal 2
1.3 The use of time-to-event data in cost-effectiveness modelling 5
1.4 Summary 8
1.5 Objectives 9
1.6 Structure of the thesis 9
1.7 Declaration 10
2 Literature review 12
2.1 Background 12
2.2 Methods 13
2.2.1 Search strategy 13
2.2.2 Eligibility 13
2.2.3 Data extraction 14
2.2.4 Synthesis 15
2.3 Results 15
2.3.1 Study selection 15
2.3.2 Study categorisation 18
2.3.3 Subset of studies generalising baseline risk 19
2.4 Discussion 21
2.4.1 Summary of approach 21
2.4.2 Summary of key findings 21
2.4.3 Interpretation of findings 21
2.4.4 Strengths and limitations 23
2.4.5 Conclusions 24
3 Time-to-event data sources and methods 25
3.1 Aggregate data 25
3.2 Individual-level data 30
3.2.1 Clinical Practice Research Datalink GOLD 30
3.2.2 CPRD GOLD lung cancer cohort 31
3.2.3 Survival analysis (Kaplan-Meier) 32
3.3 Survival analysis (parametric) 35
3.4 Cost-effectiveness modelling 35
4 Applying trial-derived treatment effects to real-world target populations: generalising cost-effectiveness estimates when modelling complex hazards 36

4.1 Background 36
4.2 Methods 37
4.2.1 Approach to model fitting 37
4.2.2 Assessing hazards 38
4.2.3 Standard parametric models 41
4.2.4 Flexible models 42
4.2.5 Model selection 44
4.2.6 Motivating example – application in cost-effectiveness modelling 45
4.3 Results 50
4.3.1 Reference treatment model 50
4.3.2 Relative effects model 54
4.3.3 Impact of generalised models on survival estimates and cost-effectiveness estimates 56
4.4 Discussion 59
4.4.1 Summary of approach 59
4.4.2 Summary of key findings 60
4.4.3 Interpretation of findings 60
4.4.4 Strengths and limitations 61
4.4.5 Conclusions 63
5 Using real-world survival estimates to inform extrapolation in Health Technology Appraisal 65
5.1 Background 65
5.2 Methods 66
5.2.1 Data sources 66
5.2.2 Approach to survival modelling 66
5.2.3 Cost-effectiveness modelling 70
5.3 Results 73
5.3.1 Study cohort 73
5.3.2 Survival analysis 73
5.3.3 Cost Effectiveness modelling 82
5.4 Discussion 83
5.4.1 Summary of approach 83
5.4.2 Summary of key findings 84
5.4.3 Interpretation of findings 85
5.4.4 Study strengths and limitations 87
5.4.5 Conclusions 89

6 Using real-world data to inform survival extrapolations: impact of censoring, sample size, and follow-up 91
6.1 Background 91
6.2 Methods 92
6.2.1 Data source 92
6.2.2 Survival analysis 94
6.3 Results 96
6.3.1 Study cohort 96
6.3.2 Survival analysis 96
6.4 Discussion 103
6.4.1 Summary of approach 103
6.4.2 Summary of key findings 103
6.4.3 Interpretation of findings 104
6.4.4 Study strengths and limitations 106
6.4.5 Conclusions 108
7 Discussion 110
7.1 Summary of key findings 110
7.2 Implications for decision-making 112
7.3 Areas for future research 115
7.4 Recommendations 118
7.5 Conclusions 119
8 Bibliography 120
Appendices 129
Chapter 2 appendix 129
Chapter 3 appendix 131
Chapter 4 appendix 135
Chapter 5 appendix 136
Chapter 6 appendix 137

List of figures

Figure 2.1. Flowchart for literature review 16
Figure 2.2. Categorisation of studies by cost-effectiveness parameter(s) generalised 19
Figure 3.1. Data type and sources used for subsequent chapters 25
Figure 3.2. Recreated Kaplan-Meier estimates for Chapter 4 observational study by van der Linden et al. (2014) 27
Figure 3.3. Overall survival from observational study from van der Linden et al. (2014) 27
Figure 3.4. Recreated Kaplan-Meier estimates for Chapter 4 RCT by Bonner et al. (2006) 28
Figure 3.5. Overall survival from Bonner et al. (2006) 28
Figure 3.6. Recreated Kaplan-Meier estimates for Chapter 5 RCT by Brahmer et al. (2015) 29
Figure 3.7. Overall survival from Brahmer et al. (2015) 29
Figure 3.8. Lung cancer mortality rates among datasets 32
Figure 3.9. June 2020 full dataset with median overall survival of 7.6 months (0.63 years) 34
Figure 3.10. June 2022 full dataset with median overall survival of 7.8 months (0.65 years) 34
Figure 3.11. CPRD GOLD and Aurum lung cancer survival estimates from Corby et al. (2024) 35
Figure 4.1. Approaches for applying trial-derived relative effects to real-world reference treatments 41
Figure 4.2. Partitioned survival model schematic with implicit transitions 47
Figure 4.3. RCT and real-world Kaplan-Meier and log-log plots 50
Figure 4.4. Selected models fitted to reference treatment. RCS: restricted cubic spline 51
Figure 4.5. Mortality probabilities for background mortality and survival modelling approaches 52
Figure 4.6. Hazard ratio plots using Cox models and restricted cubic splines with varying degrees of freedom 54
Figure 4.7. Relative effects models using restricted cubic splines 55
Figure 4.8. Cost-effectiveness modelling scenarios 57
Figure 5.1. Flowchart depicting approach to survival analysis 67
Figure 5.2. Simplified partitioned survival model structure 71
Figure 5.3. Log-cumulative hazards for age and sex strata 74
Figure 5.4 Selected survival models fitted to full real-world dataset 75
Figure 5.5. Overlaid RCT and age adjusted, sex reweighted real-world survival estimates 76
Figure 5.6. Age and sex-matched general population vs age adjusted and sex reweighted CPRD data mortality probabilities 77
Figure 5.7. Log-cumulative hazards of unadjusted, and age adjusted and sex reweighted RWD 78

Figure 5.8. Log-odds plots of unadjusted, and age adjusted and sex reweighted RWD 78
Figure 5.9. Select survival models fitted to right-censored data 80
Figure 5.10. Survival models used in cost-effectiveness model 83
Figure 6.1. Overall survival for the 60-69 stratum 96
Figure 6.2. Summary of preferred model per scenario for rMSE, RMST, and AIC 98
Figure 6.3. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=150 sample size, for each scenario 99
Figure 6.4. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=300 sample size, for each scenario 100
Figure 6.5. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=1000 sample size, for each scenario 101
Figure 6.6. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the full sample size, for each scenario 102

List of tables

Table 2.1. Studies included for classification 17
Table 4.1. Cost-effectiveness model parameters 49
Table 4.2. Formal evaluation of model fits 53
Table 4.3. Cost-effectiveness estimates (ICERs per QALY) 58
Table 5.1. Goodness-of-fit values for full real-world dataset 75
Table 5.2. Goodness-of-fit for right-censored real-world dataset via AIC and BIC values 81
Table 5.3. Accuracy of extrapolations vs full data via RMST values 81
Table 5.4. Cost-effectiveness estimates per scenario 82

1 Introduction

In this chapter, I introduce the use of cost-effectiveness analysis, real-world data (RWD), and time-to-event (TTE) data to inform decision-making in the context of Health Technology Appraisal (HTA).

1.1 The use of cost-effectiveness analysis for decision-making

Healthcare spending as a percentage of gross domestic product is set to rise at unsustainable levels for the foreseeable future.[1] To maximise population health outcomes by using increasingly scarce healthcare resources as efficiently as possible, health systems such as the National Health Service for England, among others, use HTA to inform reimbursement decisions for new treatments.[*] HTA can include assessment of both clinical- and cost-effectiveness.[2] While clinical effectiveness assessment focusses on clinical outcomes, cost-effectiveness assessment focusses on costs and benefits associated with different treatment options, often projected well beyond follow-up typically observed in clinical trials. Cost-effectiveness analysis often utilises decision-analytic modelling to estimate absolute costs and benefits with which to compare interventions. The incremental cost-effectiveness ratio (ICER), denoted as,

$$ICER = \frac{\Delta costs}{\Delta benefits} \ ,$$

uses incremental differences in costs ($\Delta costs$) and benefits ($\Delta benefits$) between health technologies or programmes to derive a summary measure of cost-effectiveness.[3] The value at which a new technology is considered cost-effective via the ICER varies by HTA body and is

[*] The National Institute for Health and Care Excellence indicates that Health Technology Appraisal falls under Health Technology Assessment, and is used for new technologies that are expected to results in additional costs for the health system.[135]

subject to conditions such as severity of disease. Costs typically include intervention and healthcare resource use costs, while benefits can be defined subject to the decision problem and consist of clinical events or holistic measures of health, such as health-related quality of life (HRQoL). In HTA, benefits are typically defined as life-years gained (LYG) or quality-adjusted life years (QALYs), which take length and HRQoL into account. The use of LYGs and QALYs allows for benefits to be compared across disease areas, which is useful for population-level decision-making. All appropriate and relevant available evidence, and underlying uncertainty, should be considered for analysis[3,4], leading to questions surrounding definitions and interpretation of sufficiently appropriate and relevant evidence.

1.2 The use of real-world data in Health Technology Appraisal

RWD is described in the National Institute for Health and Care Excellence (NICE) real-world evidence framework as patient health, experience, or care-related data collected outside of controlled clinical trial environments.[5] Guidance from The Canadian Agency for Drugs and Technologies characterises RWD as data relating to healthcare delivery or patient status from various sources, including electronic medical records, registries, administrative databases, as well as prospective sources such as pragmatic or hybrid trials.[6] The Institute for Clinical and Economic Review in the United States describes RWD as data relating to healthcare status and/or delivery collected from observations of routine clinical practice, and further includes patient-generated data via mobile devices, for example.[7] The definition of RWD thus appears broad, including routinely collected data, such as electronic health records, administrative data, and patient registries, as well as various other sources. Real-world evidence (RWE), in turn, is defined as evidence which has been generated based on analysis of RWD.[5–7] According to the NICE real-world evidence framework, uses of RWD are broadly grouped as data used to (1) contextualise clinical trials to local routine clinical practice and (2) estimate patient and system

outcomes, such as clinical outcomes or resource use.[5] These applications are aimed at better understanding and evaluating health and social care delivery in routine settings.

Currently, to assess new pharmaceutical interventions, HTA decision-making is predominantly based on evidence generated from pivotal randomised controlled trial (RCT) data used for regulatory approval. RCTs are intended to minimise issues associated with analysis of non-randomised data, such as confounding, where factors independent of the intervention can bias outcomes. There are, however, shortcomings with exclusive reliance on RCT data for HTA. Regarding the so-called efficacy-effectiveness gap[8], RCTs can lack external validity, that is the degree to which the trial findings are generalisable to a given population under investigation. This can be a problem where target populations are not adequately represented in clinical trials.[9] In contrast to trial participants, differences in age, comorbidities, and disease severity, among other characteristics, have been identified in real-world populations.[10] As decisions informed by cost-effectiveness analysis are intended to be based on estimates for target populations who will receive treatments in routine clinical practice, external validity of cost-effectiveness estimates is crucial for informed decision-making. For example, van Staa et al. (2009) claimed that the use of externally valid observational data, as opposed to clinical trial data, to quantify baseline risk would have *"radically altered the conclusions of the coxibs HTA due to the markedly different results of the cost-effectiveness analysis"*.[11] This has resulted in calls for formal assessment of external validity of cost-effectiveness estimates to be a mandatory component in HTAs.[11]

While RCTs are often designed to measure efficacy by means of a single primary endpoint, cost-effectiveness analysis draws from many parameters, in addition to efficacy, to generate cost-effectiveness estimates. Where insufficient evidence exists to adequately parameterise cost-effectiveness models, assumptions are required to fill data-gaps or project outcomes. As it is often not feasible to collect data on long-term effectiveness, safety, and costs, for example, in RCTs, RWD can be used to reduce the need for, or to inform, assumptions. Further, the strength of evidence generated from secondary or exploratory RCT endpoints to parameterise cost-effectiveness models may be limited, and one could argue that evidence generated in well-

designed real-world studies measuring the specific parameter of interest, external of the RCT, is preferred.

Interest in the use of RWD for use in decision-making is increasing due to practical benefits including reduced costs, potential for faster access to novel therapies, and novel pricing arrangements.[12] As RCTs are extremely expensive and time-consuming to conduct, where data needed to inform HTA decisions can sufficiently be captured outside of an RCT, pursuing this option is a sensible choice. Where substantial time is required to gather evidence of final outcomes, as may be the case with more chronic conditions or where lengthy trial recruitment is necessary for small patient populations, RWD can potentially be used to more rapidly generate evidence for payers, potentially resulting in expedited access to new drugs.[12] Further, where considerable uncertainty in trial-based evidence exists, RWD can be used in managed entry agreements, where new treatments are temporarily funded until additional evidence is generated.[13]

Varying degrees of acceptance regarding the use of RWD for epidemiological parameters, healthcare resource utilisation, and costs based on non-randomised data has been seen among HTA bodies in Europe.[14] A recent review on the use of RWE in NICE HTAs for oncology drugs found that HRQoL, costs, and resource use were the most common categories of RWE used in HTA submissions.[15] The review further found that criticisms surrounding the use of RWE were commonly directed at analytical methods and data sources rather than the use of RWE itself.[15]

Ongoing scepticism regarding the use of RWD in HTA is often targeted at comparative effectiveness research [16], which is interrelated with baseline risk given the requirement for estimating absolute differences in cost-effectiveness analysis. It has been noted that scepticism may stem from a lack of both literature exploring advanced methods as well as lack of trust among decision makers that findings based on RWD are robust.[16] This has been illustrated in the recent NICE consultation on the methods guidance update, where stakeholders noted a perceived resistance to the use of non-randomised data by Evidence Assessment Groups (EAG)

and Appraisal Committees acting on behalf of NICE, and requested clarity on the uses and acceptance of RWD in HTA.[17] The NICE Decision support unit (DSU)[†] Technical Support Document (TSD) 17 report has called for increased research into methods for using observational data in HTA[18], as has the NICE DSU report on RWE.[13] As the NICE real-world evidence framework[5] does not include technical approaches for integrating RWD in cost-effectiveness models as per NICE DSU documents, there appears to be an urgent need for methods guidance addressing RWD in HTA. Therefore, research including transparent evaluation and impact of methods on cost-effectiveness estimates is warranted if RWD is to be increasingly integrated in HTA processes.

1.3 The use of time-to-event data in cost-effectiveness modelling

TTE data in clinical trials are typically used to estimate relative treatment effects, commonly in the form of a hazard ratio or difference in median survival between comparators. Cost-effectiveness analysis, however, relies on absolute values and uses mean, rather than median, survival time to derive cost-effectiveness estimates.[19] Commonly, a natural history model is used in cost-effectiveness analysis to capture baseline outcomes to which relative treatment effects are applied, enabling comparison among one or more alternative treatments.[20] Therefore, appropriately capturing baseline survival, that is the probability of surviving over time (i.e. not experiencing the event of interest), to serve as a reference for comparison with another intervention is important. The use of TTE data in HTA is commonplace, with partitioned survival analysis being the most frequently used cost-effectiveness modelling approach used in oncology appraisals.[21] Extrapolation of TTE outcomes via survival modelling is required for HTA using cost-effectiveness analysis to estimate long-term outcomes, often over a lifetime time horizon.[22–25] Extrapolated estimates can have dramatic impacts on cost-effectiveness

[†] The NICE DSU is commissioned by NICE to provide technical, methodological, educational, and quality assurance advice and support to NICE. The DSU publishes assessment reports and TSDs to provide technical information on implementing approaches according to NICE's Guide to Methods of Technology Appraisal. The DSU's TSDs are widely and internationally cited as a standard for HTA methods.

estimates, often being the key driver of cost-effectiveness and thus decision-making regarding reimbursement decisions for new medicines.[26] Clinical trials typically span only one to five years[27] and are designed to measure relative effects rather than long-term survival trajectories. There can be considerable uncertainty in extrapolated estimates beyond observed clinical trial data, with individual survival model distributions often producing substantially different estimates.[26] The NICE DSU has published specific guidance for estimating relative treatment effects using RWD.[18] While longer-term external data as a valuable source of information to inform baseline risks has also been acknowledged in guidance specific to TTE data[28,29], guidance on implementation of such data for cost-effectiveness modelling appears to be lacking.

The use of external TTE data to assess plausibility of RCT-based extrapolations is currently recommended for HTA.[28] The level of detail present in available data, however, can vary significantly and may impact suitability of methods, and the data itself, for this purpose. While individual-level data including sufficient variables to allow for exhaustive adjustment is preferred, access to such data is often limited due to data sharing restrictions or the level of information included in datasets. For example, a cancer registry may include information on severity of disease and treatments, whereas a primary care dataset may not, leading to differences in possible adjustment approaches applied to datasets. Furthermore, where individual-level data are entirely unavailable, aggregate data in the form of summary plots or point estimates may be the only source of information available to analysts.[30] While such data that does not allow for exhaustive adjustment, hereafter referred to as 'limited' data, can reduce the strength of generated evidence, it is preferable to having no external data at all.[28]

A review on extrapolating RCT survival using external data by Jackson et al. (2017) has acknowledged the difficulties in quantifying uncertainty in extrapolated estimates based on RCT data alone.[27] Jackson et al. (2017) considered approaches using adjusted external data to estimate baseline risk, where extrapolated shorter-term incremental relative effects from a clinical trial can be informed by longer-term survival data from an external source. Jackson et al. (2017) acknowledge that, where external data cannot be adjusted to represent the RCT

population due to data limitations, characterising uncertainty is important, however did not include recommendations for this.[27] The NICE DSU TSD 14 has similarly recommended using adjusted external data specific to the population included in the target clinical trial to validate plausibility of extrapolated estimates.[28] The guidance, however, has cited a single methods report which was not retrievable in my search. The guidance has further stated that, in the absence of such evidence, limited external data can be used indicatively to inform plausibility of extrapolated estimates, however has provided no formal guidance on application. A critique of NICE DSU TSD 14, specific to the difficulties of interpreting the report's proposed use of external data to determine plausibility of extrapolations, has stated that candidate models for extrapolation should be tested against external data sources.[31] This critique has, in turn, been criticised for providing no advice on how to use external data for this purpose.[30] NICE DSU TSD 21, the latest guidance document on survival analysis, has acknowledged that, while Bayesian frameworks offer potential options for integrating multiple sources of evidence, survival analysis is conventionally based on fitting models to sample data in HTA.[29] The document concluded that research on the integration of external data is ongoing, and therefore has provided no firm recommendations.[29] Guidance on the actual implementation of limited RWD, with respect to access to individual-level data or included information required for extensive adjustment as previously described, thus appears to be lacking.

This absence of guidance, specifically regarding incorporation of real-world TTE baseline data into cost-effectiveness modelling, has been cited in a previous NICE HTA as a contributing factor to uncertainty in modelling approaches explored.[32] Consequently, a lack of guidance on incorporating real-world TTE data may lead to the exclusion of such data in cost-effectiveness analysis, resulting in omitted information in the evidence base used for decision-making.

Previous research has found that survival models incorporating data external to an RCT improved extrapolations.[33] This is understandable, given the relatively short time frame of RCT follow-up periods used to extrapolate outcomes over decades. Aggregate RWD has previously been used in HTA to inform potential long-term survival trends via annual survival probabilities

in terms of constant vs nonconstant survival functions for cost-effectiveness modelling.[34] Limitations on the use of RWD in this context have, however, been noted due to limited length of RWD follow-up and representativeness when RWD originates from a different country. This has led to reliance on age and sex-matched[‡] general population background mortality to assess plausibility of long-term extrapolations in some cases.[34] General population background mortality can be used to identify implausible extrapolations (i.e. where survival estimates substantially exceed those of the general population). This, however, may be considered a crude approach and has previously been critiqued as a *"weak requirement"* for survival model selection.[35]

Initiatives to understand how RWD can be used at scale across Europe in the context of HTA are currently underway, with one specific aim being the use of real-world TTE data in oncology.[36,37] A recent stakeholder workshop on integrating RWD in HTA has recommended research on methods development to pave the way for further advancements, and specifically identified natural history, such as extrapolation of cancer survival, as a key area of interest.[38] While ongoing efforts have focussed on deriving survival estimates[37], implications of and approaches for integrating RWD in decision-analytic models for HTA requires further consideration.

1.4 Summary

Substantial uncertainty regarding the use of RWD to supplement RCT data for HTA in the context of baseline TTE data exists, and research is needed to understand implications of potential approaches in this area. There appears to be particularly limited guidance with respect

[‡] For simplicity, I refer to 'age-specific' and 'sex reweighted' general population mortality/survival, where age-specific mortality/survival by sex is reweighted according to the target age and sex distribution, as 'age and sex-matched' general population background mortality/survival throughout the thesis.

to integrating limited RWD, where information needed to adjust the RWD to be representative of the target population is unavailable and no other evidence exists to inform decision-making.

1.5 Objectives

As HRQoL, costs, and healthcare resource use from RWD are already commonly used in HTA[14,15] my research aims to explore how approaches for using RWD to inform baseline survival can potentially impact decision-making in HTA. My specific objectives are:

1. To identify and categorise RWD previously used to generalise trial-based cost-effectiveness estimates in the literature and describe approaches used to integrate RWD in the context of baseline risk.

2. To investigate the impact of approaches for applying trial-derived relative effects to external, real-world baselines on cost-effectiveness estimates in the context of complex hazards and aggregate data.

3. To investigate the use of individual-level RWD, where data cannot be adequately adjusted to represent a target RCT population, to inform RCT-based survival extrapolations for HTA and assess potential impacts on decision-making.

4. To investigate the impact of censoring approaches, sample size, and follow-up time when using individual-level RWD to inform survival model selection for HTA.

1.6 Structure of the thesis

Following the introductory chapter, in Chapter 2 I present a literature review which was conducted to motivate the thesis. In the literature review, I identified cost-effectiveness modelling parameters previously used in the literature to generalise cost-effectiveness estimates

using RWD sources. The review includes an analysis of approaches used for integrating RWD in cost-effectiveness analysis in the context of baseline risk.

In Chapter 3, I provide an overview of aggregate and individual-level TTE data sources used in subsequent analytical chapters.

Chapter 4 follows from Chapter 2, in providing an overview of survival modelling approaches in the context of applying trial-derived relative effects to external, real-world baselines using aggregate TTE data. To illustrate impacts of approaches on cost-effectiveness estimates, the chapter includes reanalysis of a previously published cost-effectiveness study that aimed to generalise estimates using RWD as a motivating example.

In Chapter 5, I investigate the use of individual-level RWD to inform survival modelling approaches in HTA. The chapter includes a motivating example to demonstrate approaches and potential implications for decision-making.

Chapter 6 follows from Chapter 5, where I investigate the impact of censoring approaches, sample size, and follow-up when using RWD to inform survival modelling approaches for survival model selection.

Chapter 7 concludes the thesis with an overview of findings, implications for decision-making, areas for future research, recommendations, and overarching conclusions.

1.7 Declaration

Unless explicitly stated, all work included in the thesis is my own. I previously presented a working version of a paper based on Chapter 4 at the Health Economists' Study Group (HESG) Winter 2021 meeting, at a Health Economics Research Centre (HERC) seminar at the University of Oxford, and a paper written on the basis of Chapter 4 has been published in Value in Health, where a large part of Chapter 4 formed the published paper. I acknowledge valuable feedback received by my supervisors, members of HESG and HERC, and peer-

reviewers which has been incorporated in the chapter. I produced all components of this study independently and feedback received from co-authors is considered appropriate for a thesis. Co-authors have confirmed this in writing and given permission to re-use materials for the thesis. The publisher, Elsevier, allows articles to be included in full or in part in a thesis for non-commercial purposes and no written permission from Elsevier is necessary.

Individual-level observational TTE data is described in Chapter 3 and used in Chapters 5 and 6 of the thesis, which supplemented ongoing research conducted by the European Health Data and Evidence Network (EHDEN). Clinical Practice Research Datalink GOLD data were mapped to the Observational Medical Outcomes Partnership Common Data Model by Dr Antonella Delmestri and Ms Qai Yi Man and clinical code lists were developed and validated as part of Work Package 2 of EHDEN. I took no part in these processes and claim no credit for this work. Code for cohort generation and certain components of the survival analysis developed by Dr Danielle Newby was reused and adapted with permission for use in Chapters 5 and 6 of the thesis. I conducted all analyses independently for these chapters. Permissions for inclusion of figures from previously published studies presented in Chapter 3 are provided in captions.

2 Literature review

2.1 Background

Generalisability of cost-effectiveness estimates to target populations is important for informed reimbursement recommendations. Cost-effectiveness analysis used in HTA often uses parameters derived from pivotal clinical trials as 'points-of-truth'. However, there can be discrepancies between trial and target populations for whom decisions are being made. Differences between these populations due to strict trial inclusion criteria have been noted, raising concerns about the generalisability of outcomes observed in trials to real-world populations.[10] Understanding the impact of generalisability on cost-effectiveness estimates is therefore important. This is reflected by numerous HTA bodies' methods guidance explicitly describing generalisability or external validity as a consideration for decision-making.[22–25] Nonetheless, a lack of explicit assessment of generalisability of cost-effectiveness analyses used in HTA has previously been noted as a limitation which may have dramatic impacts on cost-effectiveness estimates and therefore reimbursement recommendations.[11]

RWD can be used to inform generalisability. However, integrating real-world parameters to supplement RCT parameters for use in cost-effectiveness modelling can be challenging. This can be due to the need for evidence on absolute estimates in cost-effectiveness analysis as opposed to evidence on relative effects often required for regulatory approval. Given cost-effectiveness estimates are driven by absolute estimates, absolute baseline risk, to which relative effects may be applied, can be important.[3] Additionally, where more complex data, such as TTE data, are used, integrating real-world and trial parameters can potentially be methodologically challenging.[3]

This chapter is aimed at identifying categories of RWD used to generalise clinical trial data for application in cost-effectiveness analysis. I therefore conducted a scoping review of the literature to identify and categorise parameters used to generalise cost-effectiveness estimates in previous studies. I focussed on studies that used RWD to generalise trial-based estimates given

the inherent use of RWD as a supplementary source of evidence for this aim. Given the importance of baseline risk for cost-effectiveness estimates and the aims of the thesis, the chapter includes a deeper analysis of approaches used in the subgroup of studies integrating real-world baseline risk.

2.2 Methods

2.2.1 Search strategy

I conducted a search through PubMed for cost-effectiveness analyses that assessed generalisability using RWD published between 01 January 2000 and 13 March 2020. Given the aim of this chapter was to identify parameters used to generalise cost-effectiveness estimates and approaches used to integrate parameters, a scoping review was deemed suitable in line with recommendations provided by Munn et al. (2018).[39] Search terms included 'cost-effectiveness analysis', 'generalisability', 'transportability', 'real-world', 'population-based', 'clinical trial', 'randomised controlled trial', and 'trial' for inclusion, and 'protocol' and 'systematic review' for exclusion in the title and abstract. The full search string, including variations, is presented in the Chapter 2 Appendix. I reviewed identified studies to ensure those already known to me through background knowledge were included in the search results. I used Rayyan software[40] to aid in abstract screening.

2.2.2 Eligibility

I included studies published between 01 January 2000 13 March 2020 that investigated generalisability of cost-effectiveness analysis findings by comparing estimates derived using trial-based vs RWD-based inputs for at least one model parameter. All populations, interventions, and comparators were considered eligible for inclusion. Duplicates were identified and removed prior to further screening. Given the aims of my review and necessary precision required to sufficiently identify individual cost-effectiveness modelling parameters, I

considered title screening to be insufficient, and undertook abstract screening for all studies identified in the search for inclusion. Studies not meeting the inclusion criteria were excluded. Specifically, publications were excluded if they were not published in the English language, if only the abstract was available, if the article type was a review, editorial, or protocol, or if the study design did not investigate generalisability of cost-effectiveness estimates substituting trial data with RWD for at least one parameter (i.e. studies including observational data alone were not included). Thereafter, I performed full text screening for publications not excluded via abstract screening to determine inclusion using the same criteria.

Similarly for categorisation, if parameters included in studies were common among scenarios, that is all scenarios used either trial or real-world parameters, they were not classified as a 'generalising' parameter. RWD was defined as any data or evidence not included or generated from a clinical trial. I further analysed the subgroup of studies applying real-world baseline risk data to identify how parameters were integrated into cost-effectiveness models.

2.2.3 Data extraction

Following extraction of included study details from Rayyan[40] into a spreadsheet, I identified and manually extracted data on cost-effectiveness model parameters used to generalise estimates per included study. I classified parameters into common categories among studies which consisted of treatment effects, baseline risks, safety/adverse events, resource use, HRQoL, discontinuation/adherence, and prescribing practices.

For the subgroup of studies using RWD to generalise baseline risk, I extracted details on the way in which treatment effects were derived (i.e. trial or real-world based) and applied to generalised baseline risks. This was done because of the interdependence between treatment effects and baseline risks for cost-effectiveness modelling. I additionally extracted details on cost-effectiveness modelling approaches used and baseline risk data type (i.e. continuous vs TTE data) for this subset of studies.

2.2.4 Synthesis

Across included studies, I summarised categorised parameters based on the frequency of usage as generalising parameters relative to the total number of studies, presented as counts and percentages. For the subset of studies including baseline risk, I presented findings as a narrative summary, given the nuances involved with application in cost-effectiveness modelling.

2.3 Results

2.3.1 Study selection

My search revealed 585 studies in total. One additional study that was known to me but not identified in the search, given it was not indexed on PubMed, was included in the review. I excluded 549 articles via abstract screening due to wrong publication/study type (n=183), no inclusion of cost-effectiveness analysis (CEA) (n=37), no comparison of trial vs RWD-based inputs (n=327), and non-English language full-text or unavailability of full-text publications (n=2). I excluded an additional 13 studies during full-text review due to no explicit trial vs RWD-based comparisons (n=9), wrong publication/study type§ (n=2), and no inclusion of CEA (n=2). This resulted in a total of 24 studies included for classification (Table 2.1). Of these, I further interpreted 14 studies including generalisation of baseline risk. A flowchart outlining study selection is depicted in Figure 2.1.

§ One publication was a proof-of-concept study and one was a review.

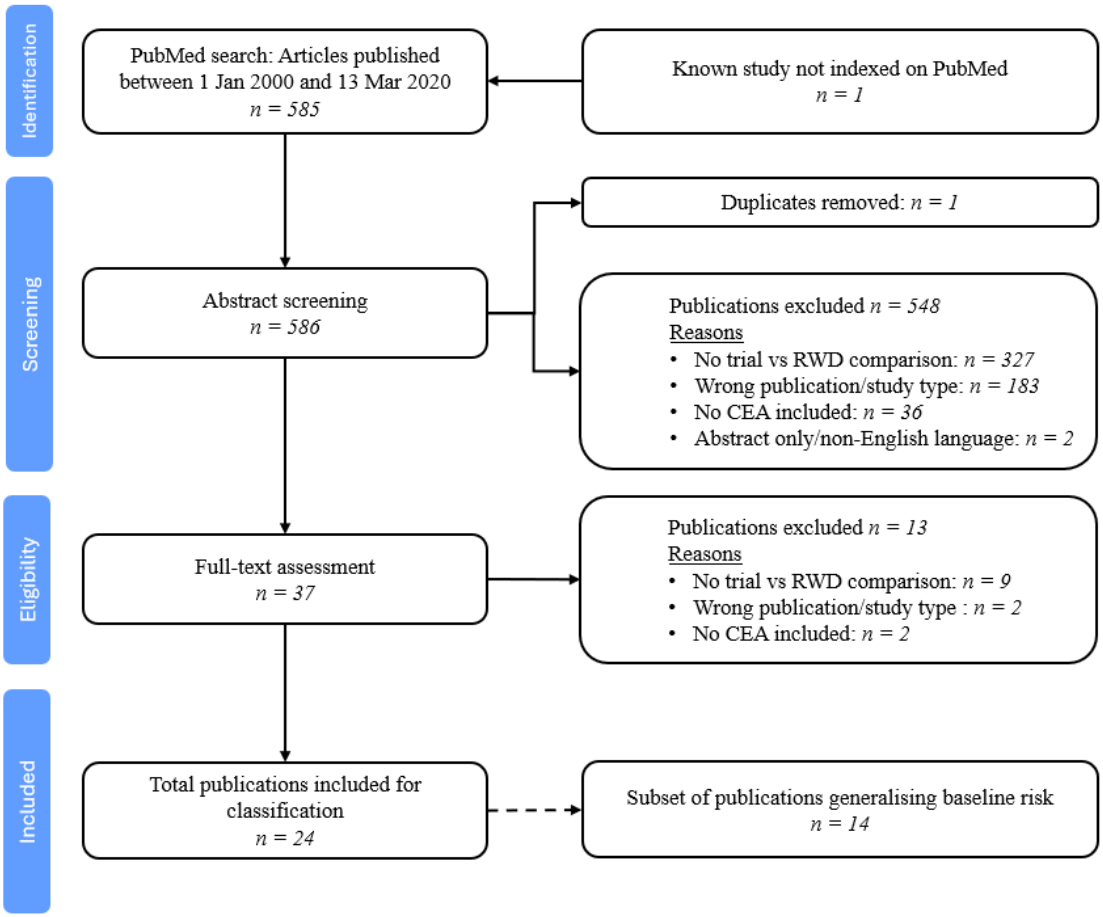


Figure 2.1. Flowchart for literature review
RWD: real-world data, CEA: cost-effectiveness analysis

Table 2.1. Studies included for classification

Study	Title	Parameters generalised
de Pouvourville et al. (2020)[41]	The contribution of real-world evidence to cost-effectiveness analysis: case study of Dabigatran etexilate in France	Treatment effects Baseline risk Safety/adverse events Resource use
Hollier-Hann et al. (2020)[42]	Updated cost-effectiveness analysis of onabotulinumtoxinA for the prevention of headache in adults with chronic migraine who have previously received three or more preventive treatments in the UK	Resource use HRQoL
de Jong et al. (2019)[43]	Cost-effectiveness of apixaban compared to other anticoagulants in patients with atrial fibrillation in the real-world and trial settings	Treatment effects Baseline risk Safety/adverse events
Desai et al. (2019)[44]	Cooled radiofrequency ablation of the genicular nerves for chronic pain due to osteoarthritis of the knee: a cost effectiveness analysis based on trial data	Resource use
Wan et al. (2018)[45]	Cost-effectiveness of continuous glucose monitoring for adults with type 1 diabetes Compared with self-monitoring of blood glucose: the DIAMOND randomized trial	Resource use
Seferina et al. (2017)[46]	Cost and cost-effectiveness of adjuvant trastuzumab in the real world setting: A study of the Southeast Netherlands Breast Cancer Consortium	Baseline risk Prescribing practices
Yeung et al. (2017)[47]	Using performance-based risk-sharing Arrangements to address uncertainty in indication-based pricing	Treatment effects Baseline risk
van Kampen et al. (2017)[48]	Real-world and trial-based cost-effectiveness analysis of bevacizumab in HER2-negative metastatic breast cancer patients: a study of the Southeast Netherlands Breast Cancer Consortium	Baseline risk
van der Linden et al. (2015)[49]	Real-world cost-effectiveness of cetuximab in locally advanced squamous cell carcinoma of the head and neck	Baseline characteristics/risk
Hartman et al. (2015)[50]	From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects	Treatment effects Baseline risk
Blommestein et al. (2014)[51]	Cost-effectiveness of rituximab as maintenance treatment for relapsed follicular lymphoma: results of a population-based study	Treatment effects Baseline risk Resource use HRQoL
Morais et al. (2014)[52]	Cost-effectiveness of rivaroxaban for stroke prevention in atrial fibrillation in the Portuguese setting	Prescribing practices
Chevallier et al. (2014)[53]	Cost-effectiveness of dabigatran versus vitamin K antagonists for the prevention of stroke in patients with atrial fibrillation: A French payer perspective	Prescribing practices
Foglia et al. (2013)[54]	Cost-utility analysis of Lopinavir/Ritonavir versus At Azanavir + Ritonavir administered as first-line therapy for the treatment of HIV infection in Italy: from randomised trial to real world	Treatment effects
van Gils et al. (2013)[55]	Real-world cost-effectiveness of oxaliplatin in stage III colon cancer: a synthesis of clinical trial and daily practice evidence	Treatment effects Baseline risk Safety/adverse events Resource use

Wouters et al. (2013)[56]	Cost-effectiveness of dabigatran etexilate in the prevention of stroke and systemic embolism in patients with atrial fibrillation in Belgium	Baseline risk
Brandes et al. (2013)[57]	Implications of real-world adherence on cost effectiveness analysis in multiple sclerosis	Discontinuation/adherence
Bellows et al. (2012)[58]	A cost-utility analysis of pregabalin versus duloxetine for the treatment of painful diabetic neuropathy	Resource use Discontinuation/adherence
Gillet et al. (2010)[59]	Delivering the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cost effectiveness analysis	Resource use
van Staa et al. (2009)[11]	A comparison of cost effectiveness using data from randomized trials or actual clinical practice: selective Cox-2 inhibitors as an example	Baseline risk Safety/adverse events Resource use
Sorensen et al. (2009)[60]	Cost-effectiveness of warfarin: Trial versus "real-world" stroke prevention in atrial fibrillation	Baseline risk Discontinuation/adherence
Annemans et al. (2008)[61]	Cost-effectiveness analysis of aprepitant in the prevention of chemotherapy-induced nausea and vomiting in Belgium	Resource use
Farahani et al. (2006)[62]	Community-based evaluation of etanercept in patients with rheumatoid arthritis	Treatment effects Baseline risk Safety/adverse events Resource use HRQoL Discontinuation/adherence
Reed et al. (2003)[63]	A Bayesian approach to aid in formulary decision making: incorporating institution-specific cost-effectiveness data with clinical trial results	Treatment effects Baseline risk Resource use

2.3.2 Study categorisation

After reviewing the publications and identifying cost-effectiveness modelling parameters generalised using RWD, I categorised parameters into the following: treatment effects, baseline risks, safety/adverse events, healthcare resource use, HRQoL, discontinuation or adherence, and prescribing practices. Figure 2.2 shows the proportion of studies including categorised parameters. As some studies included numerous generalised parameters, categories are not mutually exclusive (see Chapter 2 Appendix). The top three categories incorporated to generalise estimates were baseline risk, resource use, and treatment effects.

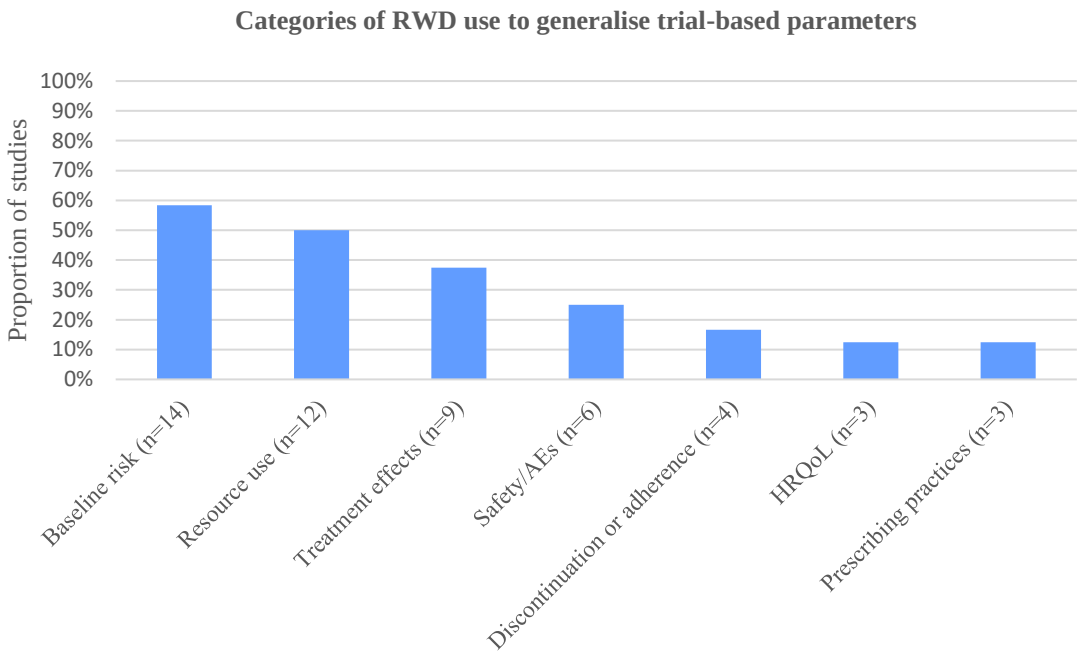


Figure 2.2. Categorisation of studies by cost-effectiveness parameter(s) generalised
RWD: real-world data, AEs: Adverse events, HRQoL: Health-related quality-of-life

2.3.3 Subset of studies generalising baseline risk

2.3.3.1 Baseline risk data type

Of the 14 studies including baseline risk to generalise estimates, six studies used TTE data involving time-dependent risks while the remaining used continuous data. Some studies using continuous data implemented time-dependent risks as appropriate via risk-adjustment and cost-effectiveness modelling (see below).

2.3.3.2 Treatment effects

Eight of the 14 studies included both real-world baseline and relative effects estimated via real-world studies.[41,43,47,50,51,55,62,63] A variety of approaches were used to derive real-world effect estimates used in cost-effectiveness modelling. Five studies used propensity score methods[41,43,50,51,63], three of which estimated relative effects using RWD alone[41,43,51], while the other two used both trial and RWD to derive relative effect estimates.[50,63] One study used regression adjustment using observational data alone to estimate relative effects.[47] One study used unadjusted real-world effect estimates and synthesised these estimates with an RCT effect

estimate for use in cost-effectiveness modelling.[55] One study used unadjusted effect estimates based on a survey study.[62] Treatment effects took the form of ratios (i.e. hazard or rate ratios) or were modelled directly via absolute differences in values between comparators.

The remaining six studies used independent relative effects taken directly from trials and applied these to real-world baseline risks for cost-effectiveness modelling. Three studies used rate ratios[11,56,60], two studies used hazard ratios[46,48],and one study calibrated trial estimates to approximate real-world estimates.[49]

2.3.3.3 Cost-effectiveness modelling

Regarding cost-effectiveness modelling approaches, of the 14 studies, nine used Markov or semi-Markov modelling approaches to derive cost-effectiveness estimates[**].[41,43,46,48,49,51,55,56,60] Two studies did not build an explicit cost-effectiveness model and used mean costs and benefits to calculate cost-effectiveness estimates.[47,62] One study used a simulation approach[11], one study used a decision-tree approach[63], and one study did not specify the approach used to derive cost-effectiveness estimates.[50]

[**] Semi-Markov approaches were used to capture varying risks over time

2.4 Discussion

2.4.1 Summary of approach

I conducted a scoping review to identify published studies aiming to generalise cost-effectiveness estimates by substituting trial-based with RWD-based cost-effectiveness modelling parameters. RWD thus acted as supplementary sources of evidence in these studies. Parameters used in studies were classified and categorised. I further analysed approaches used in studies that integrated real-world baseline risk in terms of application for cost-effectiveness modelling, including the subset of studies using TTE data for this purpose.

2.4.2 Summary of key findings

The top three cost-effectiveness analysis parameters generalised were baseline risk, resource use, and treatment effects, acknowledging that treatment effects and baseline risks were often interlinked. Most studies generalised multiple cost-effectiveness analysis parameters.

A variety of approaches were used to incorporate generalised baseline risk, including joint estimation of relative effects and baseline risk using RWD, adjustment of trial data through matching and evidence synthesis approaches, as well as application of independent relative effects from trials to real-world baselines. The majority of studies used Markov or semi-Markov approaches for cost-effectiveness modelling.

2.4.3 Interpretation of findings

Parameters generalised were selected based on the decision problem specific to each study. Therefore, my review suggests that importance and impact of individual parameter categories for decision-making based on cost-effectiveness estimates should be interpreted on a case-by-case basis. Nevertheless, baseline risk was most frequently generalised in included studies, likely due to the commonly acknowledged potential differences between trial and target real-world populations for this parameter.[10] While the majority of studies generalising baseline risk estimated baseline risk and relative effects jointly (8 of 14 studies), many (6 of 14 studies) estimated baseline risk independent of

relative effects used in cost-effectiveness modelling. This near split is of interest given the inherent limitations with both approaches. Jointly estimating baseline risk and relative effects within a single model, where baseline risk is modelled alongside relative effects by including treatment group as a covariate[28], addresses potential treatment effect heterogeneity between trial and target real-world populations. This approach additionally addresses issues surrounding covariance between parameters (see below).[64] However, treatment effect estimates using non-randomised data can suffer from limitations such as confounding, which should be explored with scenario and/or sensitivity analysis as recommended in previous research.[13] Applying trial-based relative effects to real-world baselines mitigates risk of confounding, however it assumes constant risk between trial and real-world target populations and therefore does not address potential effect heterogeneity. As touched upon earlier, appropriately quantifying uncertainty when using independent treatment effect and baseline data is problematic, given the correlation between baseline risks and treatment effects, as would be seen in a joint analysis, are not preserved.[64] While not a limitation for point estimates per se, this is a limitation for probabilistic analysis used in cost-effectiveness analysis.[64] The variation in approaches may indicate that the choice of method is dependent on data or evidence availability rather than optimal methodology.

Six studies used TTE data for parameterising baseline risk, which brings about additional challenges for cost-effectiveness modelling due to potential time-dependent risks and/or effects. When using survival modelling to extrapolate TTE data beyond observed time periods, assumptions surrounding the shape of the distributions among comparators can become key.[3] Given the importance of extrapolation for cost-effectiveness modelling, different survival distributions potentially seen in trial vs real-world populations can impact decision-making. One study calibrated trial to real-world baseline risk estimates, assuming common hazards, however this did not appear justified when examining survival data given notable differences in survival distributions between data sources. Other included studies did not provide details regarding justification of parametric models used aside from one using a simplified cost-effectiveness modelling approach based on mean values. Further, with the exception of the one study that used a simplified approach, the studies that did use TTE data

for parameterisation all used Markov modelling approaches. Markov models, however, can limit capture of time-dependent risks/events due to the inherent memoryless assumption[††]. This is a notable limitation when subsequent transitions among health states do not follow an exponential distribution, that is where risks over time are not constant. Included studies did not provide details regarding potential limitations when using a Markov approach with TTE data. Given extensive guidance on the appropriate use of TTE data for cost-effectiveness analysis for HTA[21,28,29], approaches used in the studies included in my review may therefore be limited in an HTA setting on grounds of transparency and assumptions regarding modelling approaches.

2.4.4 Strengths and limitations

Strengths of the scoping review include a systematic approach to selecting included studies, data extraction, and interpretation of findings specific to application in cost-effectiveness analysis. A broad search was conducted to appropriately identify studies, given the many nuances of cost-effectiveness modelling, and to avoid exclusion of potentially relevant studies. However, as the search was restricted to studies comparing trial vs real-world parameters, studies which aimed to generalise cost-effectiveness estimates without including reference trial estimates were excluded and may have provided further insights if included in my review.

Bias assessment was not undertaken in this review. As selection of real-world data sources used in studies was not always justified, selective use of real-world data may have potentially impacted parameters used as well as methodologies. Furthermore, possibly due to word count limitations in published papers, extensive scenario and sensitivity analyses typically required for confirmatory cost-effectiveness findings were not reported and may have provided additional information on justification for approaches used. In a similar vein, although systematic approaches to generalising cost-effectiveness models were not generally presented, this may have been considered out of scope given the aims of the papers were to assess cost-effectiveness rather than explore methods. Although a relatively recent review of HTAs using RWD for treatment effect estimation concluded with similar

[††] The studies that did use semi-Markov models to account for time dependent risks did not use TTE data to estimate transition probabilities for cost-effectiveness modelling

findings, in terms of insufficient transparency and lack of scenario and sensitivity analyses surrounding real-world inputs[13], an updated review of previous HTAs may have provided additional insights. Further, a review of HTAs considering RWD beyond treatment effects may have offered additional perspectives on approaches used and how these were perceived. Further, I conducted my search using PubMed only, and searching via other databases such as the University of York's Centre for Reviews and Dissemination (CRD) database[65], the International Network of Agencies for Health Technology Assessment Database (INAHTA)[66], and Tufts Medical Center's Centre for the Evaluation of Value and Risk in Health (CEVR) Cost-Effectiveness Analysis (CEA) registry[67] may have identified additional publications that were not identified in my search. Lastly, considering the numerous nuances involved with cost-effectiveness modelling, it was not feasible to exhaustively report on modelling approaches or the potential interdependencies between generalised parameters. Furthermore, given the number of assumptions involved and lack of detail provided in some studies, it is possible that I may have misinterpreted some approaches. [67]

2.4.5 Conclusions

In conclusion, among the studies included in my review, a range of parameters for generalising cost-effectiveness estimates using RWD were identified. To generalise estimates, RWD most often replaced trial data for baseline risk, followed by resource use and treatment effects. My review highlights the considerable variation in parameters used and/or required for generalising cost-effectiveness estimates. It further illustrates that methodological guidance for integrating real-world baseline risk data, especially in the case of TTE data, may be warranted, given the variation in methods and lack of systematic approaches seen in included studies.

3 Time-to-event data sources and methods‡‡

In this chapter, I describe aggregate and individual-level observational data used for survival analysis, as well as common methods included in the analytical chapters of the thesis. Figure 3.1 illustrates data sources and methods for aggregate and individual-level data included in Chapters 4, 5, and 6. Data sources used for other cost-effectiveness modelling parameters, such as costs and utilities, as well as specific parametric survival modelling approaches employed, are presented in individual chapters.

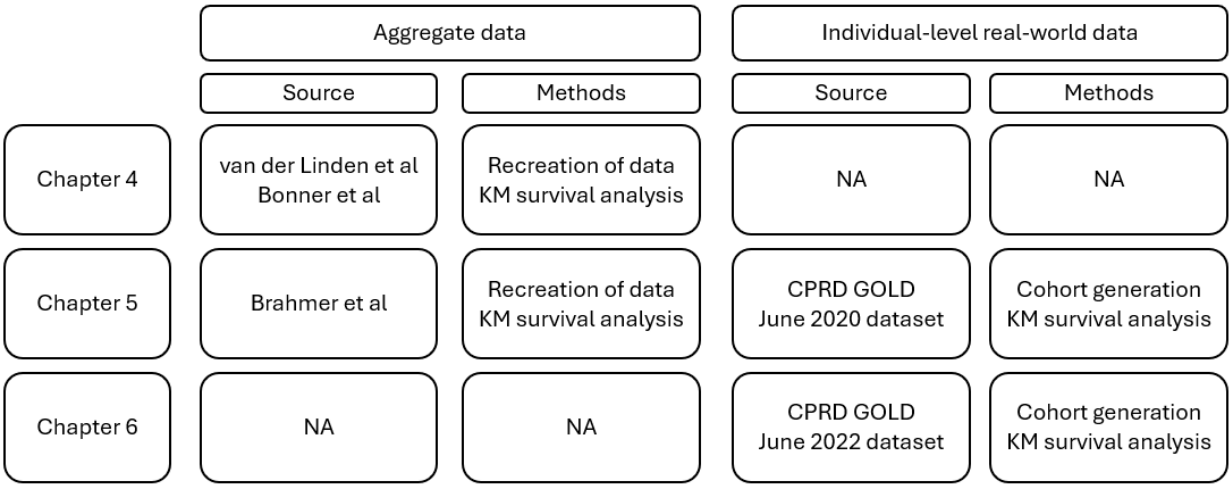


Figure 3.1. Data type and sources used for subsequent chapters

KM: Kaplan-Meier, NA: not applicable, CPRD GOLD: Clinical Practice Research Datalink GOLD dataset

3.1 Aggregate data

In Chapter 4, I explore the use of aggregate data alone for application of survival analysis approaches in cost-effectiveness modelling. I used a previously published study identified in Chapter 2 that aimed to generalise a cost-effectiveness analysis by using reference treatment risks based on an observational study, among other parameters, in place of RCT reference treatment risk.[49] The

‡‡ With reference to Section 3.2 - *Individual-level data*, my contributions as a researcher to the specific deliverables of EHDEN Work Package 2 are not included in the thesis. Clinical Practice Research Datalink GOLD data were mapped to the Observational Medical Outcomes Partnership Common Data Model by Dr Antonella Delmestri and Ms Qai Yi Man and clinical code lists were developed and validated as part of Work Package 2 of EHDEN. I took no part in these processes and claim no credit for this work. Code for cohort generation and certain components of the survival analysis developed by Dr Danielle Newby was reused and adapted with permission for use in Chapters 5 and 6 of the thesis. I conducted all analyses independently for these chapters.

observational study investigated generalisability of overall survival outcomes from an RCT by estimating overall survival using data from the Dutch National Cancer Registry and comparing outcomes.[68] The RCT investigated cetuximab plus radiotherapy vs radiotherapy alone in locally advanced squamous cell carcinoma of the head and neck.[69] Given only summary data were available via published Kaplan-Meier (KM) plots, I recreated TTE data for analyses.

For recreation, I digitised all KM estimates using WebPlotDigitizer[70] and extracted numbers at risk presented in the RCT publication. As risk tables accompanying KM estimates were unavailable in the observational study publication, numbers at risk were estimated as per the approach previously described by Jansen (2011)[71] as

$$n_t = \frac{n_{t+1}}{S_{t+1}/S_t},$$

where n_t is the total number at risk calculated per time interval, taking the numbers censored identified via plots between intervals into account. I used 10-month intervals and the total number included in the observational study at $n_{t=0}$ for calculations.

I deemed this necessary as initial recreation without numbers at risk led to invalid KM estimates, in-line with previous findings that estimates can be inaccurate in the absence of numbers at risk used in the recreation algorithm.[72] I manually assessed validity of extracted point survival estimates to ensure values approximated published KM curves. I identified implausible individual digitised values (i.e. increasing survival over time) via error handling and adjusted these manually. Specifically, where I found that digitised survival percentages increased over time, they were manually inspected vs published KM plots and manually overwritten to conform to published KM curves. Following cleaning of digitised data, I created pseudo-individual-level data using the algorithm developed by Guyot et al. (2012).[72] KM estimates were then produced and visually compared vs the original publications for validity (see Figure 3.2 to Figure 3.5). Figure 3.4 and Figure 3.5 show a slight discrepancy in median survival between recreated and published estimates for the radiotherapy alone arm (30.1 and 29.3 months median survival for recreated and published estimates, respectively), potentially due to the 'steps' in the KM curves being slightly different as a result of the recreation.

However, given the underlying data were used for fitting parametric survival models, I deemed recreated data sufficient for subsequent analysis.

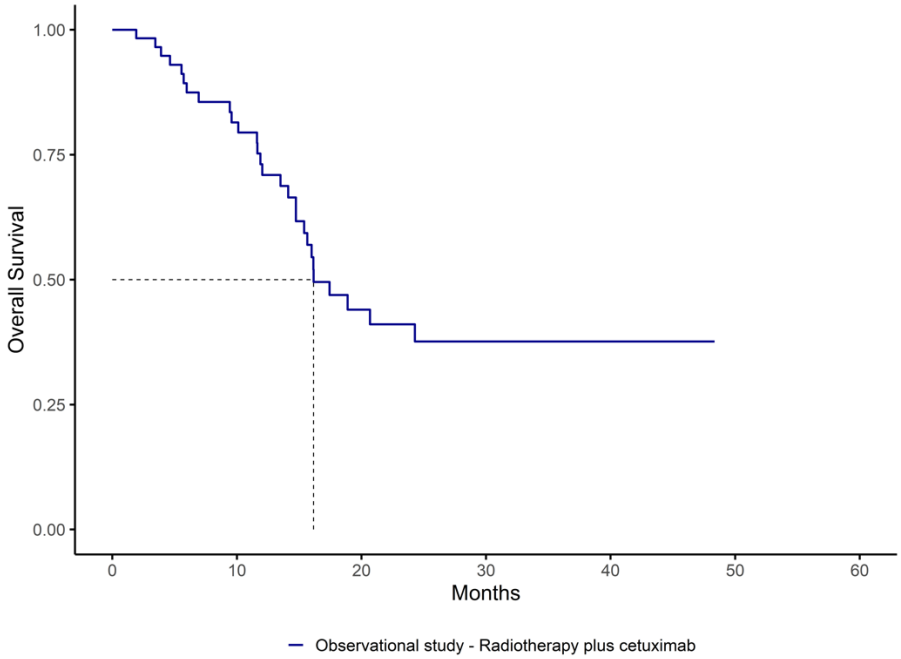


Figure 3.2. Recreated Kaplan-Meier estimates for Chapter 4 observational study by van der Linden et al. (2014)

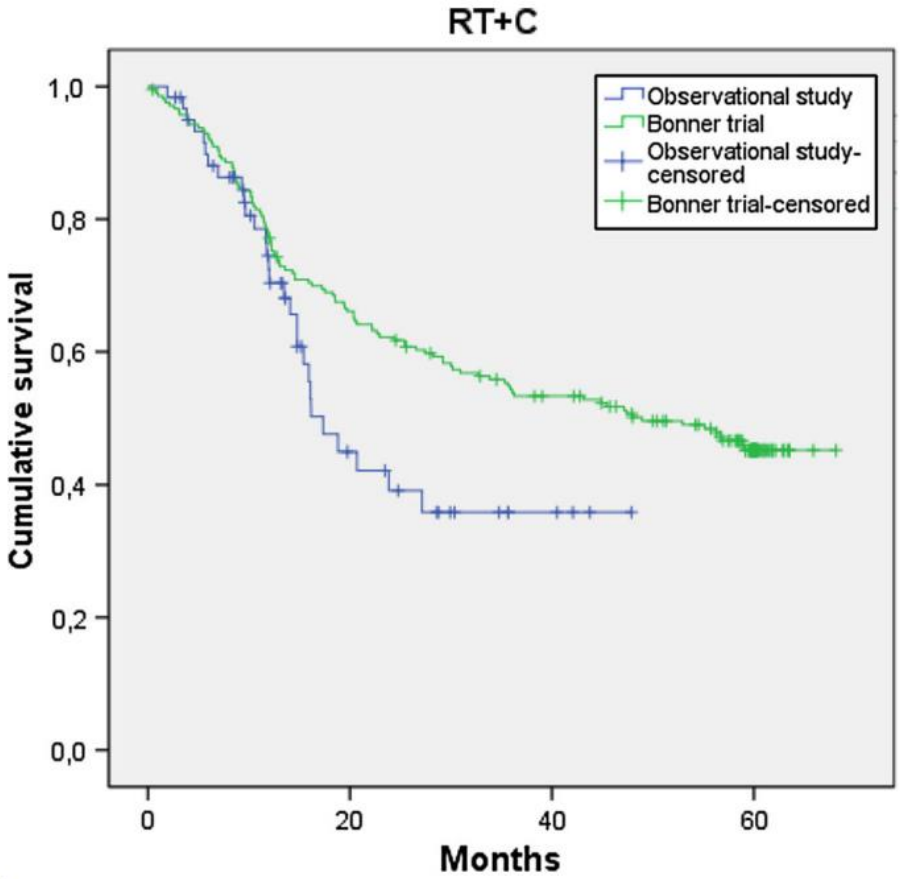


Figure 3.3. Overall survival from observational study from van der Linden et al. (2014)

Permission for use in Thesis received from SpringerNature, 23 May 2024. License number: 5794620437329

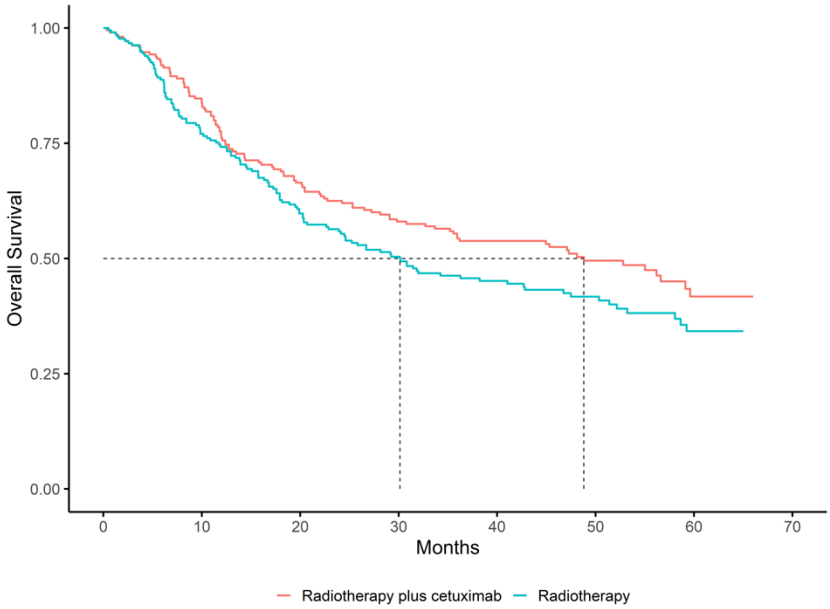


Figure 3.4. Recreated Kaplan-Meier estimates for Chapter 4 RCT by Bonner et al. (2006)

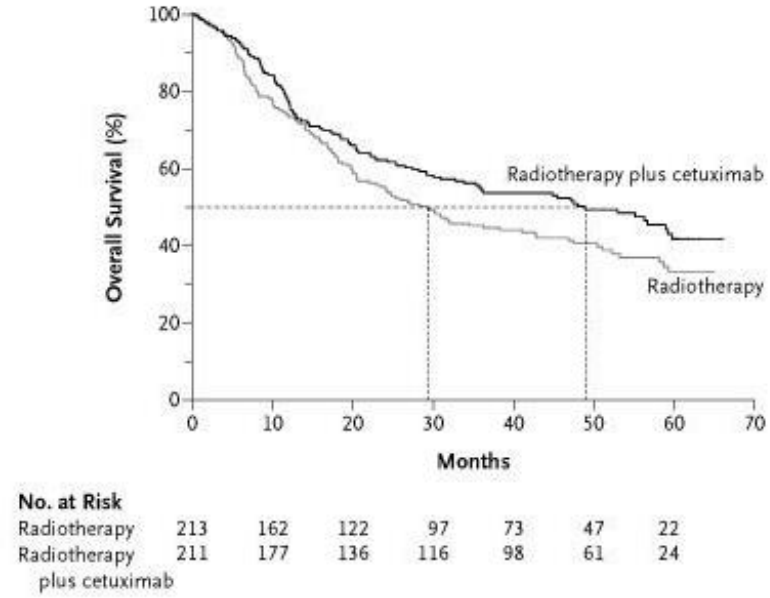


Figure 3.5. Overall survival from Bonner et al. (2006)

Reproduced with permission from Bonner JA, Harari PM, Giralt J, et al. *Radiotherapy plus Cetuximab for Squamous-Cell Carcinoma of the Head and Neck.* New England Journal of Medicine *2006; **354**: 567–78, Copyright Massachusetts Medical Society.*

Chapter 5 incorporates aggregate data (alongside individual-level observational data described in the following section) to allow for replication of survival modelling approaches used in the motivating example, which was based on a previous HTA. I created pseudo-individual data following the

approach used for Chapter 4[§§] using aggregate data from an RCT comparing Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer (see Figure 3.6 and Figure 3.7).[73]

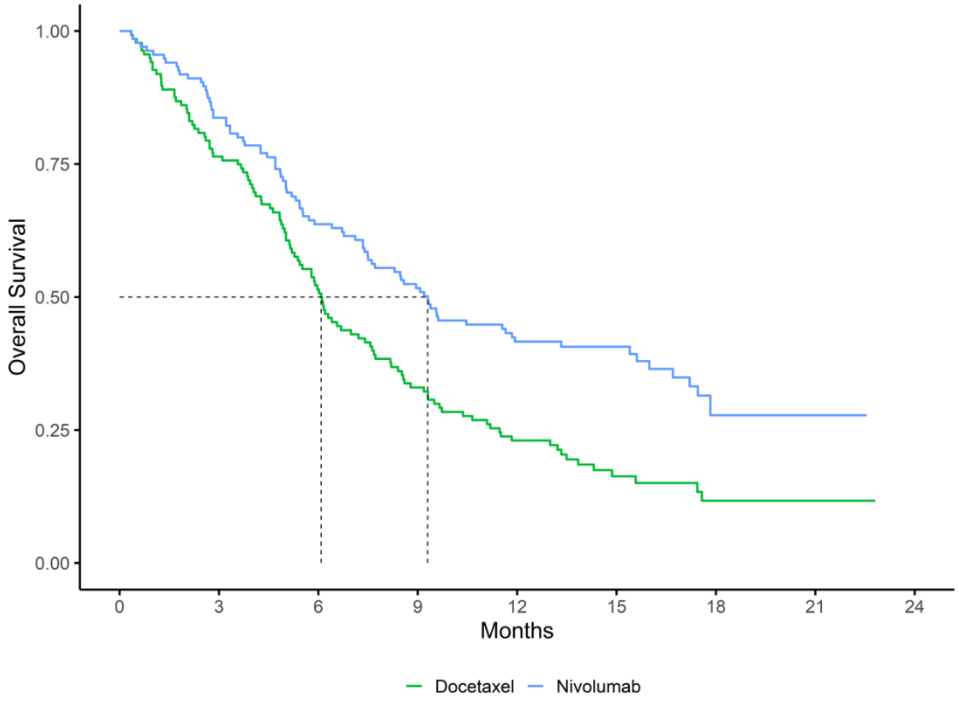


Figure 3.6. Recreated Kaplan-Meier estimates for Chapter 5 RCT by Brahmer et al. (2015)

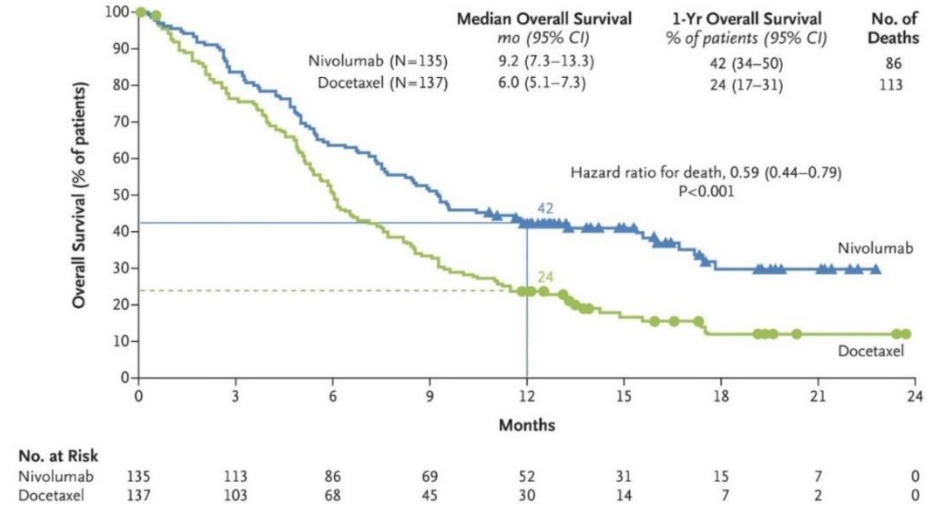


Figure 3.7. Overall survival from Brahmer et al. (2015)

Reproduced with permission from Brahmer J, Reckamp KL, Baas P, et al. *Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer.* New England Journal of Medicine *2015; **373**: 123–35, Copyright Massachusetts Medical Society.*

[§§] Given risk tables were included in the source publication, numbers at risk were not approximated as per the approach described for Chapter 4.

3.2 Individual-level data

3.2.1 Clinical Practice Research Datalink GOLD

The work conducted in Chapters 5 and 6 supplemented ongoing research by the European Health Data and Evidence Network (EHDEN) on the feasibility of generating RWE for HTA purposes in oncology (Work Package 2).[36] Chapter 5 explores practical application of real-world TTE data for cost-effectiveness modelling using individual-level RWD from Clinical Practice Research Datalink (CPRD) GOLD. Chapter 6 examines the impact of censoring, length of follow-up, and sample size when using RWD to inform survival model selection for cost-effectiveness analysis is investigated.

CPRD is a primary care database containing medical records from general practitioner (GP) practices in the UK, and its use in observational research has been substantial.[74] CPRD contains pseudonymised individual-level information on demographics, lifestyle data, clinical diagnoses, prescriptions, and preventive care. CPRD has ethics approval from the Health Research Authority to support research using pseudonymised patient data and has permission to conduct studies without consent under Section 251 of the NHS Act 2006. Individual-level data used in my studies were obtained through an approved application to the CPRD (application number 22_001843). Only summary results are presented throughout the thesis, ensuring individual anonymity is maintained.

Chapters 5 and 6 include analyses using June 2020 and June 2022 CPRD GOLD datasets[75,76], respectively. In my analyses, CPRD GOLD data which had been mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) were used.[77,78] Briefly, the OMOP CDM standardises the structure of observational data, allowing analyses to be conducted in a consistent and reproducible manner. Native information contained within observational datasets (such as CPRD GOLD) must first be mapped to standardised tables within the OMOP CDM. Once mapped, software tools designed for use with the CDM can be used, thereby making analyses agnostic to the source dataset used given relevant fields are available.

3.2.2 CPRD GOLD lung cancer cohort

In Chapters 5 and 6, I conducted real-world TTE data analyses using lung cancer cohorts generated from the previously described CPRD GOLD datasets. For both CPRD GOLD datasets, lung cancer-specific overall survival was estimated using a retrospective, observational cohort study design. Identical criteria were used to generate cohorts for both datasets. Individuals with a lung cancer diagnosis between 1 January 2000 and 1 January 2019 who were aged 18 years or older with at least 365 days of prior history and information on age and sex were included in both studies. Individuals were followed up until either exit from the database, date of death, or end of the study period (31 December 2019) and censored if no event (death) had been recorded.

I used Observational Health Data Sciences and Informatics (OHDSI) tools for analyses, including the ATLAS webAPI, and CohortDiagnostics, CDMConnector, and CohortGenerator R packages to generate the lung cancer cohorts.[79–82] Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) diagnostic codes were used to identify lung cancer diagnoses. Diagnostic codes indicative of either non-malignant cancer or metastasis, or melanoma and lymphoma were excluded. The clinical code list used to identify the lung cancer cohort was developed as part of the broader EHDEN work package[36] described in Section 3.2.1 and is presented in the Chapter 3 Appendix (Table A.3.1). The code list was validated by clinicians as part of the broader work package. In CPRD GOLD, previous research has shown that the majority (84%) of recorded lung cancer diagnostic codes consist of "Malignant tumour of lung" (44%) or "Primary malignant neoplasm of lung" (40%)[83] (see Chapter 3 appendix). Stratification by lung cancer subtype was therefore not feasible, and implications of this are discussed in Chapter 5.

The CPRD GOLD datasets were not linked to other datasets. Linkage to other datasets may have provided additional information regarding classification of the cohort and timely recording of events (such as cancer registry data, in particular). However, this was not possible due to time constraints and the logistics involved with datasets approved for use in the broader work package previously described.[36] I therefore considered potential implications of using unlinked data to generate survival estimates. Previous research investigating discrepancies in lung cancer mortality rates using linked

data vs CPRD GOLD alone has shown that diagnosis dates were delayed using CPRD GOLD[84], which in turn would impact survival estimates. However, these differences were relatively small (Figure 3.8), and I therefore expected survival estimates based on CPRD alone to be sufficient for the purposes of Chapters 5 and 6 given the importance of hazards trends, rather than absolute estimates, used to inform approaches for cost-effectiveness modelling.

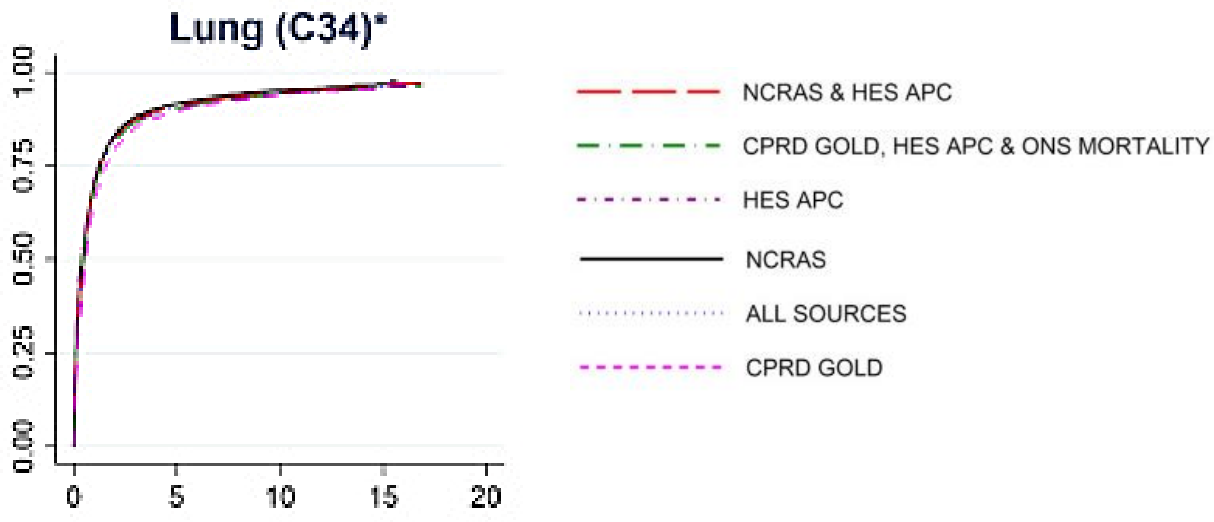


Figure 3.8. Lung cancer mortality rates among datasets

Figure adapted from Strongman et al. (2020). to include only lung cohort. Copyright © Strongman H, Williams H, and Bhaskaran 2020. Re-use permitted under CC BY. Published by BMJ.

3.2.3 Survival analysis (Kaplan-Meier)

For KM estimation using individual-level CPRD data in Chapters 5 and 6, mortality was defined as all-cause mortality based on recorded date of death and used to estimate overall survival. For my analyses, KM estimates assuming last-observation carried forward were firstly used to validate survival estimates. For validation, I compared KM estimates against a recently published study by van Veelen et al. (2022) which assessed 12 subgroups of lung cancer using CPRD GOLD, with each subgroup stratified by differing inclusion criteria to understand survival estimates in populations considered eligible vs ineligible for selected RCTs.[85]

Using visual inspection, the study by van Veelen et al. (2022) showed similar survival trends compared with my estimates using CPRD GOLD and among the various subgroups identified in their

study. van Veelen et al. (2022) estimated median survival ranges spanning 6.6 to 10.8 months across subgroups, which includes the 7.6 and 7.8 median survival estimates from my analyses. For all subgroups presented by van Veelen et al. (2022), on crude visual inspection, KM curves showed evidence of common survival trends, indicating limited heterogeneity of survival distributions observed via KM estimates.

I additionally compared my survival estimates with another study by Corby et al. (2024) that used a modified inclusion criteria. The study excluded individuals with a prior history of another cancer diagnosis using both CPRD GOLD and Aurum[***].[86] Visual inspection of survival distributions and interpretation of median survival demonstrated largely consistent estimates between the different inclusion criteria. Median survival estimates presented in the study by Corby et al. (2024) were 8.5 and 7.9 months for Aurum and GOLD, respectively.[86] While slightly higher, these estimates are relatively consistent with my median estimates of 7.6 and 7.8 months for the June 2020 and June 2022 CPRD GOLD datasets. KM survival estimates for both CPRD GOLD datasets are presented in Figure 3.9 and Figure 3.10, and alongside the external study in Figure 3.11.

In Chapter 5, I use real-world lung cancer survival to inform modelling approaches using a previous HTA as a motivating example. This entailed making the real-world dataset as representative as possible of the population represented in the decision problem of the HTA. The lack of data on staging, cancer type, and other typical trial inclusion criteria included in CPRD limited adjustment approaches to age and sex. However, previous research has shown that 80-85% of lung cancer patients are diagnosed with non-small cell lung cancer[87], with 64-71% being classified as having stage 3 or 4 lung cancer.[88] Given the non-small cell stage 3 or 4 lung cancer inclusion criteria for the HTA target population[73], I expect the CPRD GOLD lung cancer cohort used in Chapter 5 to be broadly representative. Relatedly, while CPRD GOLD has previously been found to be broadly representative of the UK population[74], with decreasing numbers of English GP practices included in the database in recent years[75,76], the impact of decreasing representativeness of CPRD GOLD on estimates is

[***] CPRD GOLD uses Vision® software whereas Aurum uses EMIS Web® software and are released as separate databases due to differences in structure and coding between the systems. CPRD GOLD includes practices across the UK whereas CPRD Aurum includes practices in England only.[76,136,137]

uncertain. Nonetheless, agreement in estimates between CPRD GOLD and Aurum from Corby et al. (2024)[86] (Figure 3.11) provides assurance that estimates are valid for the purposes of Chapters 5 and 6.

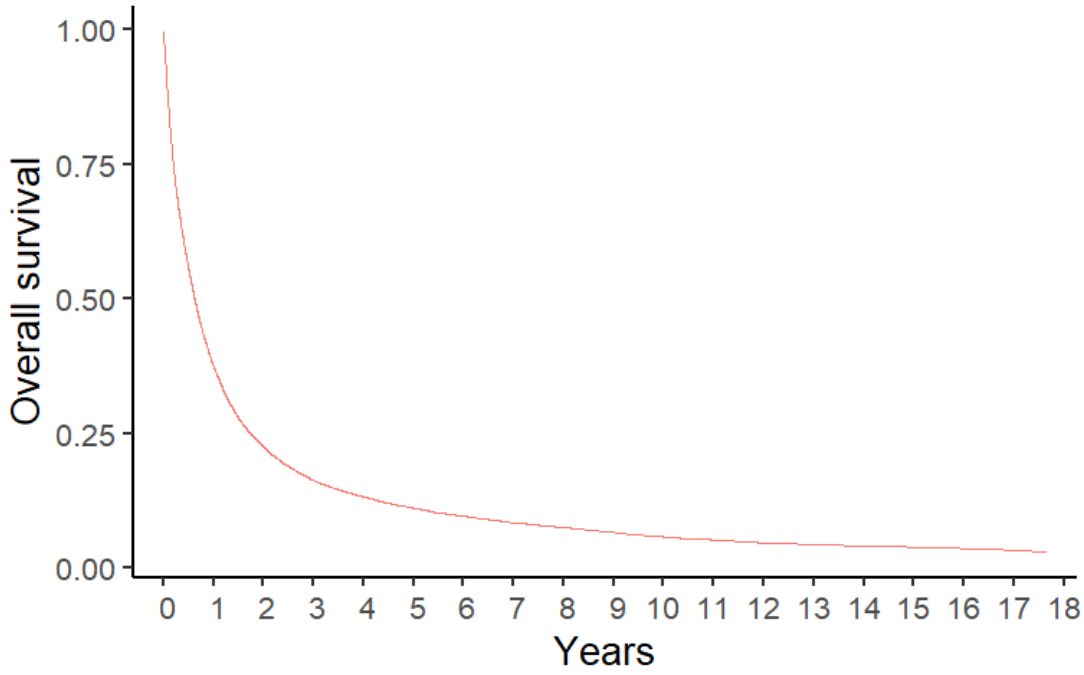


Figure 3.9. June 2020 full dataset with median overall survival of 7.6 months (0.63 years)

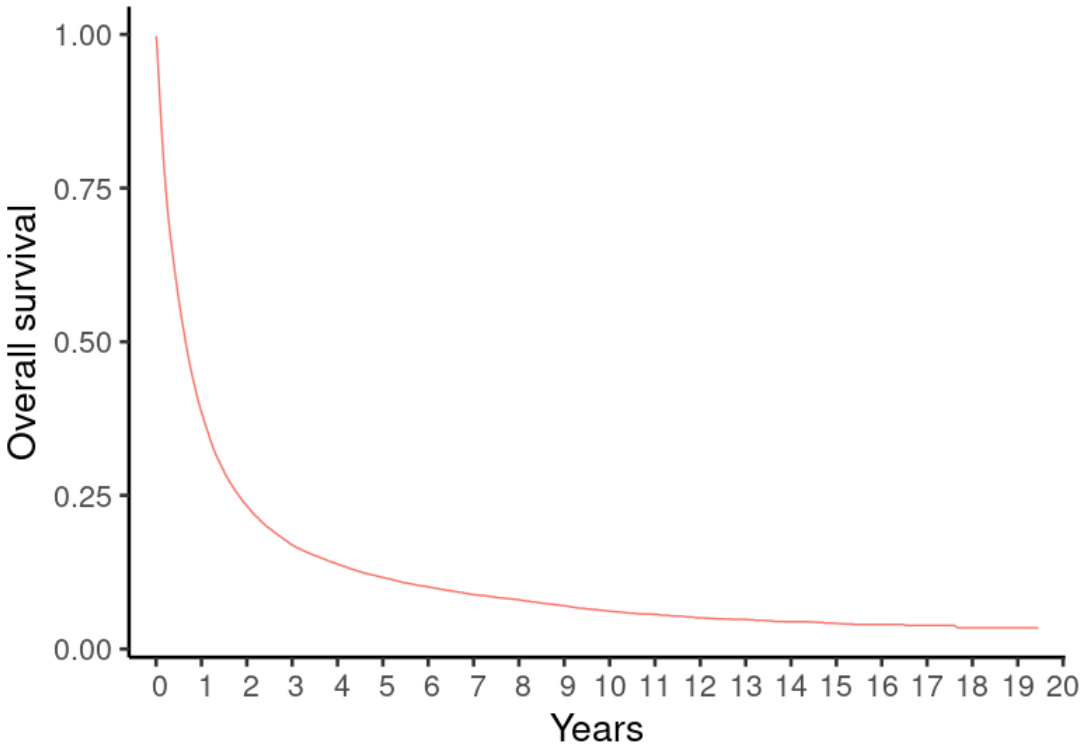


Figure 3.10. June 2022 full dataset with median overall survival of 7.8 months (0.65 years)

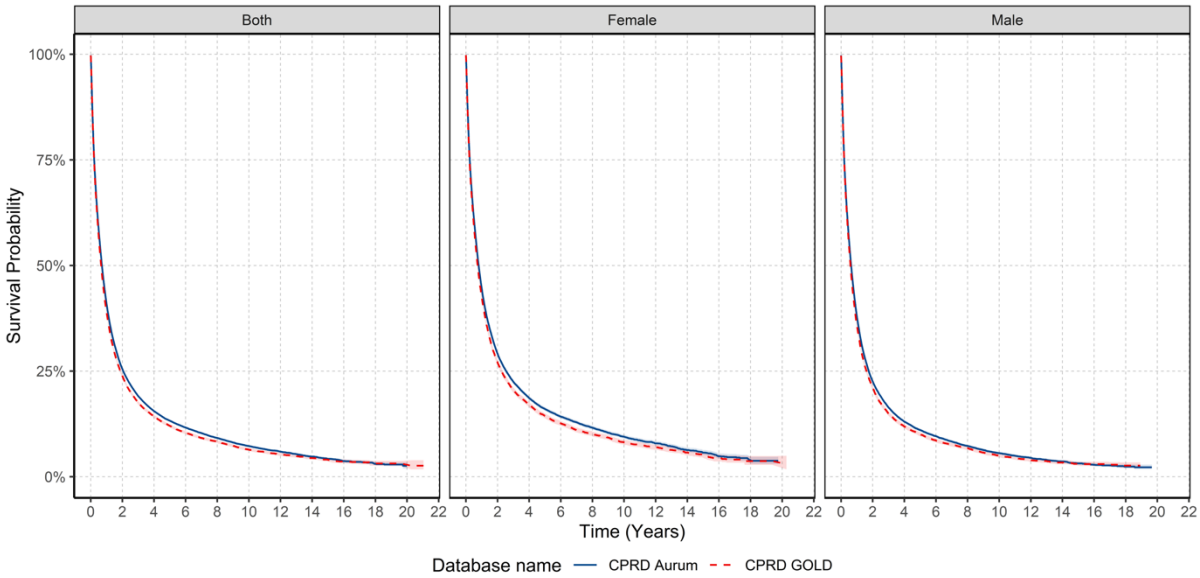


Figure 3.11. CPRD GOLD and Aurum lung cancer survival estimates from Corby et al. (2024)

Permission from senior author received for use in thesis.

3.3 Survival analysis (parametric)

Following validation performed via assessment of KM estimates for Chapters 4, 5, and 6, parametric models were fitted using individual or recreated pseudo-individual-level data. I followed NICE DSU TSD 14[28] guidance on survival analysis for cost-effectiveness modelling[28] as appropriate, and described all analyses in respective chapters. I performed all survival analyses using R.[89] Specifically, the 'survival'[90] and 'flexsurv'[91] packages were used for all survival analyses and 'tidyverse'[92] packages were used for producing plots. The 'survminer'[93] package was additionally used to produce some survival plots.

3.4 Cost-effectiveness modelling

I performed all cost-effectiveness modelling in Chapters 4 and 5 in base R[89] using independently written bespoke code. The 'tidyverse' packages were used for some data processing.[92] Extreme value checks were used to ensure validity of cost-effectiveness results. Specific details for cost-effectiveness modelling are presented in respective chapters.

4 Applying trial-derived treatment effects to real-world target populations: generalising cost-effectiveness estimates when modelling complex hazards[†††]

4.1 Background

As described in Chapter 2, generalisability of trial-based cost-effectiveness estimates to real-world target populations is an important consideration in HTA. Where new therapies are not yet available, as can be the case during HTA when reimbursement recommendations are being decided, relative effect estimates must necessarily be restricted to those observed in clinical trials.[28] Moreover, access to individual-level data for relative effect and/or baseline risk estimation may be limited, leading to dependence on aggregate data for parameterising cost-effectiveness models.[30] In the context of independent aggregate TTE baseline risk and relative effects data, complex hazards, where hazards change over time for either or both baseline and treatment effect estimates, can make modelling of data for use in economic evaluation challenging.

A number of previous studies identified in Chapter 2 assessed generalisability of trial-based cost-effectiveness estimates to real-world populations using real-world baselines with trial-derived relative effects applied. In the context of TTE data, combining baseline hazards with relative effects from external sources can bring about challenges when handling data that includes complex hazards and

[†††] Parts of the work presented in this chapter have been published in and presented as:

Koblbauer I, Prieto-Alhambra D, Burn E, Pinedo-Villaneuva R. Applying trial-derived treatment effects to real-world populations: generalizing cost-effectiveness estimates when modelling complex hazards. *Value Health*. 2024;27(2):173-181.

Koblbauer I. Applying trial-derived treatment effects to real-world populations: how to estimate cost-effectiveness when modelling complex hazards. HERC Talks and Events. 08 Sept 2022

Koblbauer I, Prieto-Alhambra D, Pinedo-Villanueva R. Application of trial-derived relative effects to real-world baselines for cost-effectiveness analysis. HESG Winter meeting 2021.

The publisher, Elsevier, allows articles to be included in full or in part in a thesis for non-commercial purposes and no written permission from Elsevier is necessary. This chapter includes additional background and analysis that was not included in published or presented work.

time-varying relative effects. Applying relative effects to real-world baselines can thus be methodologically challenging when trying to best capture data for parameterising cost-effectiveness models. As noted in Chapter 2, previous investigation of approaches for appropriately applying trial-derived relative effects to independent, real-world target populations in the context of complex hazards appears limited.

In this chapter, I investigate methods which can be used to apply trial-derived relative treatment effects to external real-world baseline TTE data when faced with complex hazards using aggregate data. Following a description of proposed methods for analysing TTE data, the chapter continues with a motivating example to illustrate potential impacts of approaches on cost-effectiveness estimates.

4.2 Methods

4.2.1 Approach to model fitting

In this chapter, I aimed to provide guidance on generalising cost-effectiveness estimates when applying trial-derived relative effects to independent real-world baselines in the context of complex hazards. I focussed on approaches for fitting survival models to sample data (i.e. trial or real-world population samples) and did not include survival modelling using a Bayesian framework. Given baseline survival can refer to one of two (or more) treatment arms, I hereafter refer to this as reference treatment survival for this chapter. I first provide an overview of assessing hazards to inform approaches as well as an overview standard and selected flexible survival modelling approaches for applying trial-derived relative effects to external reference treatments. I then apply approaches in a motivating example to illustrate approaches and potential implications for decision-making.

Regarding TTE data used, I assumed that no external evidence existed to supplement trial-derived relative effects regarding potential treatment effect heterogeneity between trial and target populations and that trial-derived relative effects were applicable to the real-world population. I therefore applied

relative effects as observed in the RCT. Assuming that observed KM estimates were valid, curve fitting was aimed at capturing observed hazards.

4.2.2 Assessing hazards

To assess hazards, visual inspection of log-log plots can be used to understand proportionality of hazards between the trial and real-world reference therapy and trial relative effects. Current guidance states that log-log plots can be used to assess proportionality of treatment effects to inform whether the proportional hazards assumption is violated, however other approaches including Schoenfeld residual plots can additionally be used.[28] Proportionality of treatment effects are particularly important for cost-effectiveness modelling, given the potential importance of extrapolated estimates on cost-effectiveness outcomes. Hazard ratio plots can be used to further assess time-varying effects using kernel-based[94] or spline-based models.[91] Log-log plots, specifically log-cumulative hazards plots, can be visually inspected to assess whether hazards between comparators or independent data sources are parallel. If lines remain parallel throughout the observed period, proportional hazards can be assumed, and the use of a hazard ratio applied to the reference survival distribution justified. Where hazards cross, the use of a hazard ratio is inappropriate given the proportional hazards assumption is violated, and survival distributions should be modelled independently to capture time-varying effects.[28] Where survival distributions are unavailable for both comparators, and modelling relies on relative effects, flexible approaches for applying time-varying effects can be used. Where survival distributions do not cross yet are not parallel, scenarios evaluating constant and time-varying effects can be explored to understand potential impacts on cost-effectiveness estimates. Nevertheless, as survival distributions are not typically expected to be completely parallel, visual inspection remains subjective. Care must therefore be taken when interpreting log-log plots to inform approaches for capturing treatment effects in cost-effectiveness modelling. Further cost-effectiveness modelling strategies, such as tapering of treatment effect beyond the observed period, or application of common hazards after a specified timepoint, can be used to inform potential impact of uncertainty in modelled estimates.

Log-log plots can additionally be used to inform modelling approaches for capturing reference hazard functions, where non-linear log-log plots indicate distinct hazards segments which may need to be captured using flexible survival models. While log-log plots are commonly evaluated on the log-cumulative hazards scale to evaluate suitability of various parametric models through inclusion/exclusion, various specifications can be used to evaluate suitability of particular parametric models.[95]

To evaluate exponential and Weibull distributions, log-time vs log-cumulative hazards, denoted as,

$$\log(-\log(S(t))) ,$$

where $S(t)$ is the survivor function, can be used. Given the Weibull distribution can include the exponential distribution as a special case dependent on the value of the shape parameter (i.e. if the slope of the log-cumulative hazard plot approximates unity an exponential distribution may be suitable). To evaluate Gompertz distributions, time vs log-cumulative hazards can be used since the log-hazard function is linear to time on the natural scale. To evaluate loglogistic models, log-time vs log-odds, denoted as,

$$\log\left(\frac{S(t)}{1-S(t)}\right) ,$$

can be used. To evaluate lognormal models, log-time vs the standardised normal distribution, denoted as,

$$\Phi^{-1}(1- S(t)) ,$$

where Φ is the standard normal cumulative distribution, can be used. Given the lognormal distribution is very similar to the loglogistic distribution[28,95], evaluation of log-log plots on either scale may be sufficient for determining appropriateness of distributions for cost-effectiveness modelling.

Given the generalised gamma distribution can include the Weibull or lognormal distribution as special cases, dependent on the value of the scale parameter, log-log plots used for these distributions can be used to inform suitability. In practice, while suitability of parametric models can be informed by

visual inspection of log-log plots, evaluation of formal goodness-of-fit, visual inspection on the natural scale, and plausibility of extrapolation should all be used together to inform preferred model selection[28], making the use of log-log plots one of numerous criteria used to aid in survival modelling for HTA.

I refer to NICE DSU TSD 14[28] and TSD 21[29] for a broader overview of survival analysis methods and guidance for survival analysis in the context of HTA, which I followed as appropriate. NICE DSU TSD 14 provides guidance on fitting survival models using RCT data in the context of cost-effectiveness analysis for HTA and is internationally referenced by HTA bodies regarding methods for curve fitting. NICE DSU TSD 21 provides guidance on the use of flexible survival models where evidence of complex hazards exists.

Proportionality between trial and real-world reference treatments, and of treatment effects, would make the use of hazard ratios to capture treatment effects and/or to calibrate trial to real-world reference treatments appropriate. However, nonproportional hazards indicate that the use of flexible approaches may be suitable to capture distinct hazards segments for reference treatments and relative effects.

In order to compartmentalise strategies for appropriately modelling hazards when using independent data sources for real-world reference treatments and trial-derived relative effects, Figure 4.1 illustrates a simplified structure to guide approaches. Given the data used in this chapter, I concentrated on the bottom-right scenario in the context of complex hazards. I additionally explored scenarios assuming proportionality to understand sensitivity of estimates to approaches.

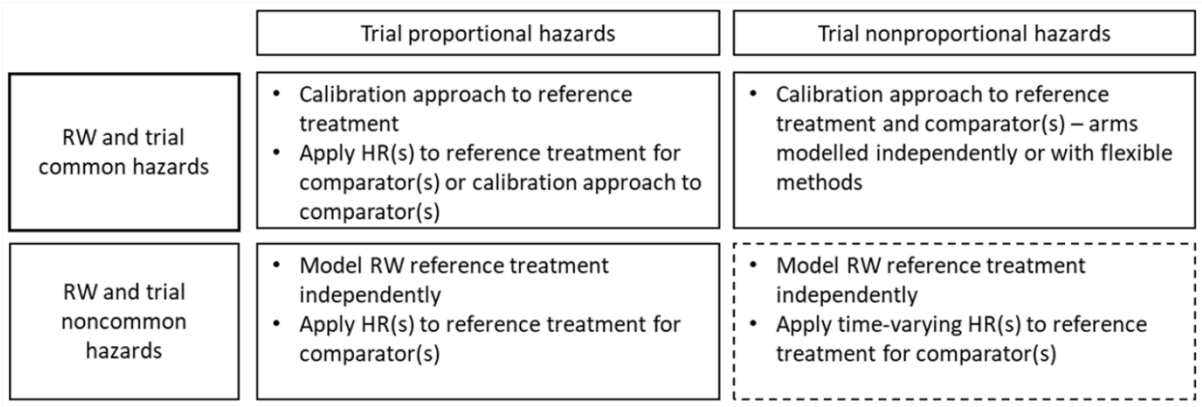


Figure 4.1. Approaches for applying trial-derived relative effects to real-world reference treatments
RW: real-world, HR: hazard ratio.

4.2.3 Standard parametric models

Standard parametric models fit hazards as linear functions of time or log time and include proportional hazards and accelerated failure time models. Standard parametric models refer to exponential, Weibull, Gompertz, lognormal, loglogistic, and generalised gamma distributions as per those listed and recommended for use in HTA.[28] Using the Weibull distribution on the log-cumulative hazards scale (the log survival function – Equation (4.1)) as an example, it can be seen from Equation (4.2) that the shape parameter γ is a function of log time $ln(t)$ with the intercept $ln(\lambda)$.

$$S(t) = \ln H(t) \quad [4.1]$$

$$\ln H(t) = \ln \lambda + \gamma \ln(t) \quad [4.2]$$

For relative effects, as shown in Equation (4.3), adding a covariate $\chi\beta$ creates a proportional hazards model, that is, that relative effects are not dependent on time.

$$\ln H(t) = \ln H(t) + \chi\beta = \ln \lambda + \gamma \ln(t) + \chi\beta \quad [4.3]$$

In the case of multi-parameter distributions such as the Weibull, relative effects can also be placed on additional parameters (e.g. the shape γ parameter in addition to the scale λ parameter) to better handle nonproportional hazards.[28,96] This application being successful, however, is dependent on the given distributions adequately fitting the reference treatment TTE data.

4.2.4 Flexible models

4.2.4.1 Spline-based models

Cubic splines are characterised as a set of piecewise polynomials joined at smoothed points (referred to as knots). Cubic spline functions thus model hazards using spline rather than linear functions of log time, allowing for more flexibility over standard parametric functions.[97] Restricted cubic splines (RCS) restrict splines to be linear at the tails, making them applicable for survival modelling where extrapolation is needed. Complex hazards seen in reference treatment models, as well as time varying relative effects, can thus be better captured using spline-based vs standard models.[98] Using an RCS on the hazards scale[99], it can be seen by Equation (4.4) that by fitting multiple spline functions between knots, a range of shapes can be captured.

$$lnH(t) = \lambda_0 + \lambda_1 Z_1 + \lambda_2 Z_2 + \ldots + \lambda_{m-1} Z_{m-1} \quad , \qquad [4.4]$$

where *m* is the total number of knots (including boundary knots) and Equations (4.5-4.7) are the basis spline functions Z_j calculated as

$$Z_1 = ln(t) \qquad [4.5]$$

$$Z_{j-1} = \left(Z_1 - k_j\right)^3 - \tau_j (Z_1 - k_{min})^3 - \left(1 - \tau_j\right)(Z_1 - k_{max})^3 \quad , \qquad [4.6]$$

where k_j are the internal knot locations, k_{min} and k_{max} are the boundary knot locations, and τ_j is calculated as

$$\tau_j = \left(k_{max} - k_j\right)/(k_{max} - k_{min}). \qquad [4.7]$$

Boundary knots are placed at minimum k_{min} and maximum k_{max} uncensored log event times. Internal knots are commonly located in evenly distributed quantiles of estimated uncensored log event times.[99] Where two independent data sources are used for the reference treatment and time-varying relative effects models, as can be the case when generalising trial-based relative effects to real-world reference treatments, common knot locations of log time are needed among data sets in order to apply time-varying treatment effects appropriately.[100] Although previous research has shown that fits are

relatively insensitive to knot placement[98], sensitivity to knot placement can be tested, with placement informed by clinical expert opinion or external evidence. Current guidance, however, recommends using default knot placement, given selective knot placement can be highly subjective.[101] The number of knots should be informed by formal goodness-of-fit, visual inspection, and plausibility of extrapolation (as per standard parametric models). Sample size and number of events in the dataset should be considered when selecting the number of knots, and sensitivity analysis is recommended to understand the impact of varying the number of knots.[29]

Using Equation (4.4), treatment effects can be placed on the intercept $\lambda_{j=0}$ for a proportional hazards model, as well as on higher order terms λ_j, where $j > 0$, to capture time-varying relative effects. Time-varying treatment effect estimates can be evaluated for sensitivity to placing treatment effects on higher-order spline terms. Placement of effects on particular higher-order terms with treatment effect degrees of freedom > 1 can be informed via visual inspection of log-log plots of the trial data. Again, placement of effects on selective knots can, however, be highly subjective.

4.2.4.2 Piecewise models

An alternative flexible approach, the use of piecewise models, was popularised following a publication by Bagust and Beale (2014) which challenged NICE TSD 14 guidance on approaches for survival analysis used in HTA.[30] Briefly, Bagust and Beale (2014) noted practical limitations with the published guidance, including restricted access to individual-level data, and reliance on log-cumulative hazard plots and proportional hazards models. NICE TSD 14 recommended standard survival models and assessment of plausibility of extrapolations were further critiqued, and an alternative framework for survival analysis and extrapolation of TTE data was presented, focussing on the use of piecewise models. Piecewise models make use of distinct hazards segments identified in TTE data which are modelled independently, typically using standard parametric distributions. An example of this being a piecewise constant hazards model, which uses exponential distributions with segments conditional on distinct time intervals, as represented in Equation (4.8),

$$h(t) = \begin{Bmatrix} \lambda_0 \;\; if & t \leq \theta_0 \\ \lambda_1 \;\; if & \theta_0 < t \leq \theta_1 \\ \vdots & \vdots \\ \lambda_i \;\; if & \theta_i < t \leq \theta_{i+1} \end{Bmatrix} \;, \qquad [4.8]$$

where λ_i are the distinct shape parameters corresponding to segments defined by time θ_i. Inspection of log-log plots can be used to identify distinct hazards within TTE data, both within a baseline distribution and between comparators. Although segment selection, with regards to time, is necessarily subjective, fitting models to distinct segments allows for a variety of shapes and relative effects to be captured. As segments identified between the baseline and relative effects model may differ, a compromise in terms of segment specification may be required in order to achieve common segments for both models. Numbers of observations should be considered when selecting segments, given low numbers can lead to issues with curve fitting. Further, care must be taken when selecting segments, given estimates can be sensitive to time points chosen.[19]

4.2.5 Model selection

Reference treatment and relative effects models can be assessed for goodness-of-fit via the Akaike information criteria (AIC) and the Schwartz, or Bayesian, information criteria (BIC).[102,103] Both the AIC and BIC use the maximum likelihood estimator (MLE) and number of model parameters to estimate fit (see Equations 4.9 and 4.10),

$$\text{AIC} = -2\ln(L) + 2k \;, \qquad [4.9]$$

$$\text{BIC} = -2\ln(L) + \ln(N)k \;, \qquad [4.10]$$

where L is the log-likelihood function for the MLE, N the total number at risk, and k the number of parameters in the model.

Visual inspection of goodness-of-fit can be assessed against trial and real-world KM curves. Plausibility of extrapolations should be assessed via longer term data and/or expert clinical opinion[28] although can often be restricted to age and sex-matched background mortality due to data limitations.

4.2.6 Motivating example – application in cost-effectiveness modelling

As a motivating example, I applied approaches in various scenarios drawing on overall survival data used in a previous study assessing the generalisability of a trial-based cost-effectiveness analysis of radiotherapy plus cetuximab (RT+C) versus radiotherapy (RT) alone in non-squamous cell carcinoma of the head and neck.[49,68,69] The study showed a variety of hazards and time-varying relative effects among the trial and RWD, making it an appealing example for illustrating methods. The previous study conducted a cost-effectiveness analysis using observed survival data directly (without fitting parametric survivals for extrapolation), using a basic approach to calibrate RCT to RWD without directly assessing the impact of the real-world reference treatment distribution. Application of parametric survival modelling approaches in a cost-effectiveness model were thus deemed important to illustrate the potential impacts of various survival modelling approaches on cost-effectiveness estimates, and in turn decision-making.

As described in Chapter 3, all overall survival TTE data were digitised and recreated using the algorithm presented by Guyot et al. (2012)[72] to derive the real-world baseline model and trial-based relative effects.[68] Recreated estimates were validated against source estimates. Standard parametric (exponential, Weibull, Gompertz, lognormal, loglogistic, and generalised gamma), piecewise exponential, and RCS models were fitted to the overall survival data. RCS models with varying numbers of internal knots and degrees of freedom for treatment effects were fitted on the log cumulative hazards scale. I refer to Rutherford et al. (2015)[98] for considerations on selecting a sufficient number of knots for spline models, which I followed given the potential for using unnecessarily complex models. Briefly, Rutherford et al. (2015) performed a simulation study assessing restricted cubic splines modelled on the log-cumulative hazards scale for capturing complex hazards and found that using more than 2 degrees of freedom was generally sufficient for capturing complex hazards.[98] Rutherford et al. (2015) further argue that care must be taken to avoid overly complex models due to the risk of overfitting. Although assessment of formal goodness-of-fit may indicate complex models are favoured, especially when using large datasets, less complex models may achieve very similar fits and therefore be preferred on grounds of parsimony.[98]

To assess the impact of different approaches in-light of previous literature on the subject, I followed an approach used in the previously published study by van der Linden et al (2015).[49] In this study, trial TTE data were calibrated to approximate real-world reference treatment TTE data, implicitly assuming common hazards. To best replicate this approach, hereafter referred to as the calibration scenario, I estimated a hazard ratio via survival probabilities between data sources to calibrate RCT survival to real-world median survival. This, effectively, shifted the RCT distributions (fitted independently per arm to preserve all relative effects) to approximate median survival observed in the real-world study. I used the RCS model which best fitted the RCT data for this purpose given the use of KM curves alone would not allow for extrapolation.

As previously noted, for the purposes of exploring approaches, I assumed that all KM estimates were valid, in that the aim was to best capture observed time-varying relative effects and hazards in the reference distribution through survival modelling approaches. As RWD were presented for the RT+C group alone in the referenced studies, I followed this as the reference treatment in the current study. As I focussed on methodological approaches, rationale for the use of RT+C as the reference treatment is beyond the scope of this chapter and can be found in the reference studies.[49,68]

I developed a basic three-state partitioned survival cost-effectiveness model to estimate cost-effectiveness via the ICER per QALY to understand potential impacts of approaches on decision-making (Figure 4.2). I chose a partitioned survival modelling approach for numerous reasons. First, partitioned survival models use survival estimates directly and do not rely on strong assumptions used when transforming data to generate transition probabilities, therefore maintaining simplicity while preserving precision in estimates used in cost-effectiveness modelling. Second, the use of state-transition models, such as Markov models, when using TTE data can be problematic, in that subsequent transitions between health states (e.g. from ‘progressed’ to ‘dead’ as illustrated in Figure 4.2), require either an assumption of constant hazards or individual-level data to derive transition probabilities. These transition probabilities can be volatile depending on timing of individual events observed in the individual-level data. Given individual-level data were not available and a constant hazards assumption did not appear valid based on the observed data, a Markov modelling approach

was deemed inappropriate for my study. An overview on the use of partitioned survival models vs other modelling approaches for use in HTA is provided in NICE DSU TSD 19.[21]

Given the methodological rather than confirmatory nature of the study, I focussed on differences in, and direction of results as opposed to confirmation of cost-effectiveness. Sensitivity analyses (i.e. probabilistic analysis and one-way sensitivity analysis) for survival estimates as well as the various other inputs populating the cost-effectiveness model, as would be required for HTA, were not included.

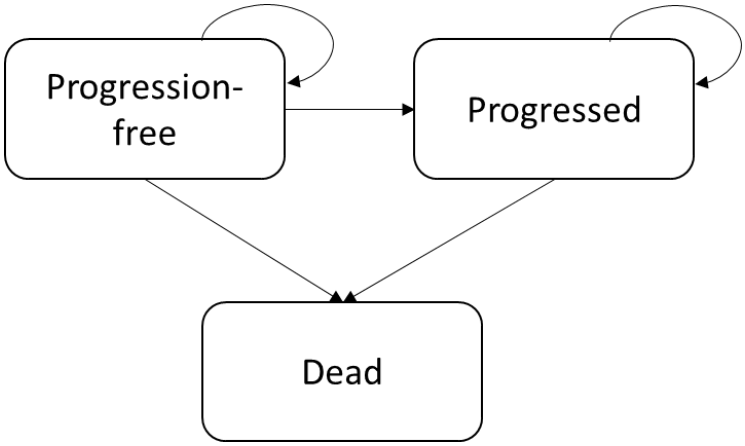


Figure 4.2. Partitioned survival model schematic with implicit transitions

Briefly, the model included 'progression-free', 'progressed' (progressed disease), and 'dead' health states. These health states were based on common endpoints used in clinical trials and precedent cost-effectiveness modelling approaches used in HTA.[21] The 'progression-free' health state was defined as the area-under-the-curve (AUC) of the progression-free survival distribution used in the cost-effectiveness model. The 'dead' health state was defined as the complement of the AUC of the overall survival distribution. The 'progressed' health state was defined as the difference in AUC between the overall survival and progression-free survival distributions. As progression-free survival KM estimates were not available in sufficient detail for recreation, a surrogacy relationship with overall survival was assumed. Progression-free survival was thus approximated using overall survival calibrated to median progression-free survival (i.e. common hazards were used as a simplifying assumption). This assumption was based on previous research investigating surrogacy relationships which found strong correlations between event-free survival and overall survival in patients with

locally advanced head and neck cancer.[104,105] Costs, utilities, and other parameters were taken from the original publication[49] (Table 4.1), with 'progressed' costs approximated given insufficient detail provided in the published study. For costs, currency as per the original publication (euros) was applied uninflated in the model. ICERs are presented in both euros as well as converted pound sterling for ease of interpretation. 'Progression-free' and 'progressed' costs and utilities were artificially capped when background mortality was appended. I applied this cap with the assumption that ongoing, constant costs over the duration of the lifetime time horizon did not seem plausible. I used this assumption to adopt a conservative approach for estimating differential costs and utilities between comparators. I used a lifetime time horizon to capture costs and benefits, where < 1% of the cohort were still alive at the end of the time horizon for all scenarios. Annual discount rates were taken from the original publication, which followed Dutch guidance.[106]

Appendage of background mortality, often referred to as a 'cure point', was performed using age and sex-matched background mortality.[107] 'Cure point' time was informed by the 5-year assumption made by van der Linden et al. (2015)[49] and inspection of mortality probabilities between extrapolations and general population background mortality. Background mortality was appended in the cost-effectiveness model using transition probabilities calculated as per Equation (4.11)

$$tp(t_i) = \frac{S_{LT}(t_i)}{S_{LT}(t_{i-1})}, \qquad [4.11]$$

where $tp(t_i)$ is the transition probability and $S_{LT}(t)$ the survivor function computed via life tables at time interval i. The transition from modelled to background survival at the 'cure point' was applied as per Equation (4.12)

$$P(t) = \left\{ \begin{array}{r} S(t) \;\; if \;\; t < m \\ S(t-1)[tp(t)] \;\; if \;\; t = m \\ P(t-1)[tp(t)] \;\; if \;\; t > m \end{array} \right\}, \qquad [4.12]$$

where m represents the time at which the 'cure point' was applied and $P(t)$ the survival probability used in the cost-effectiveness model.

Background mortality was taken from life tables[107] for the calendar year corresponding to the real-world study and the target population country (the Netherlands).

Table 4.1. Cost-effectiveness model parameters

Cost-effectiveness analysis model parameters

Parameter	Input	Source
Progression-free mean costs per month		
Years 0 to 2	€597.47	
Year 3 to 'cure point'	€57.00	
Progressed mean costs per month		
Year 0 to 'cure point'	€2,702.81	
Progression-free utility		
Year 0 to time horizon	0.86	van der Linden et al. (2015)[49]
Progressed utility		
Year 0 to 'cure point' (PFS utility applied thereafter)	0.28	
Discounting		
Costs	0.040	
Benefits	0.015	
Time horizon	Lifetime*	NICE health technology evaluations: the manual. Process and methods. 2022[25]
PFS HR	0.70	Bonner et al. (2006)[69], van der Linden et al. (2015)[49]
PFS surrogacy HR (applied to RT+C OS)	1.93	Bonner et al. (2006)[69], NICE TA145 ERG report[108]

PFS: Progression-free survival, HR: hazard ratio, ERG: Evidence Review Group (now Evidence Assessment Group)

* Defined as sufficient time to capture survival or benefits that last for the remainder of a person's life. In all scenarios < 1% of the cohort were still alive at the end of the time horizon

4.3 Results

4.3.1 Reference treatment model

Visual inspection of log-log plots (Figure 4.3 right panels) revealed evidence of complex hazards for both trial relative effects and real-world reference treatment data, indicating that evaluation of flexible modelling approaches was warranted. Non-parallel hazards, specifically crossing of cumulative hazards between trial and real-world reference treatments (Figure 4.3 bottom right panel) showed strong evidence of violation of the proportional hazards assumption, indicating that a common model to approximate the real-world reference treatment via a 'calibration' hazard ratio may have been insufficient. Further, non-linear hazards seen in the real-world reference treatment indicated that flexible survival models would likely be required.

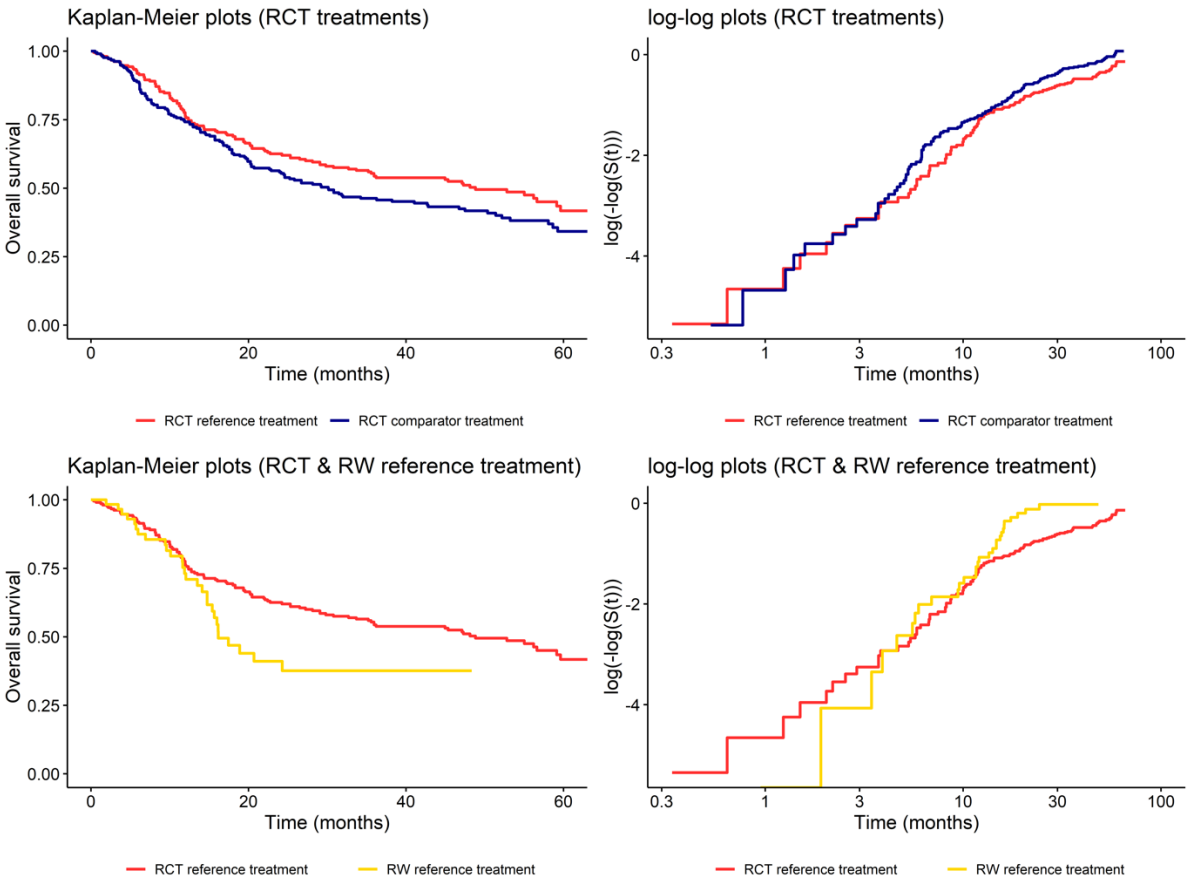


Figure 4.3. RCT and real-world Kaplan-Meier and log-log plots

RCT: randomised controlled trial, RW: real-world

Standard parametric models fitted to the real-world TTE data showed poor capture of hazards in terms of visual inspection and formal evaluation via AIC and BIC values (Figure 4.4 and Table 4.2).

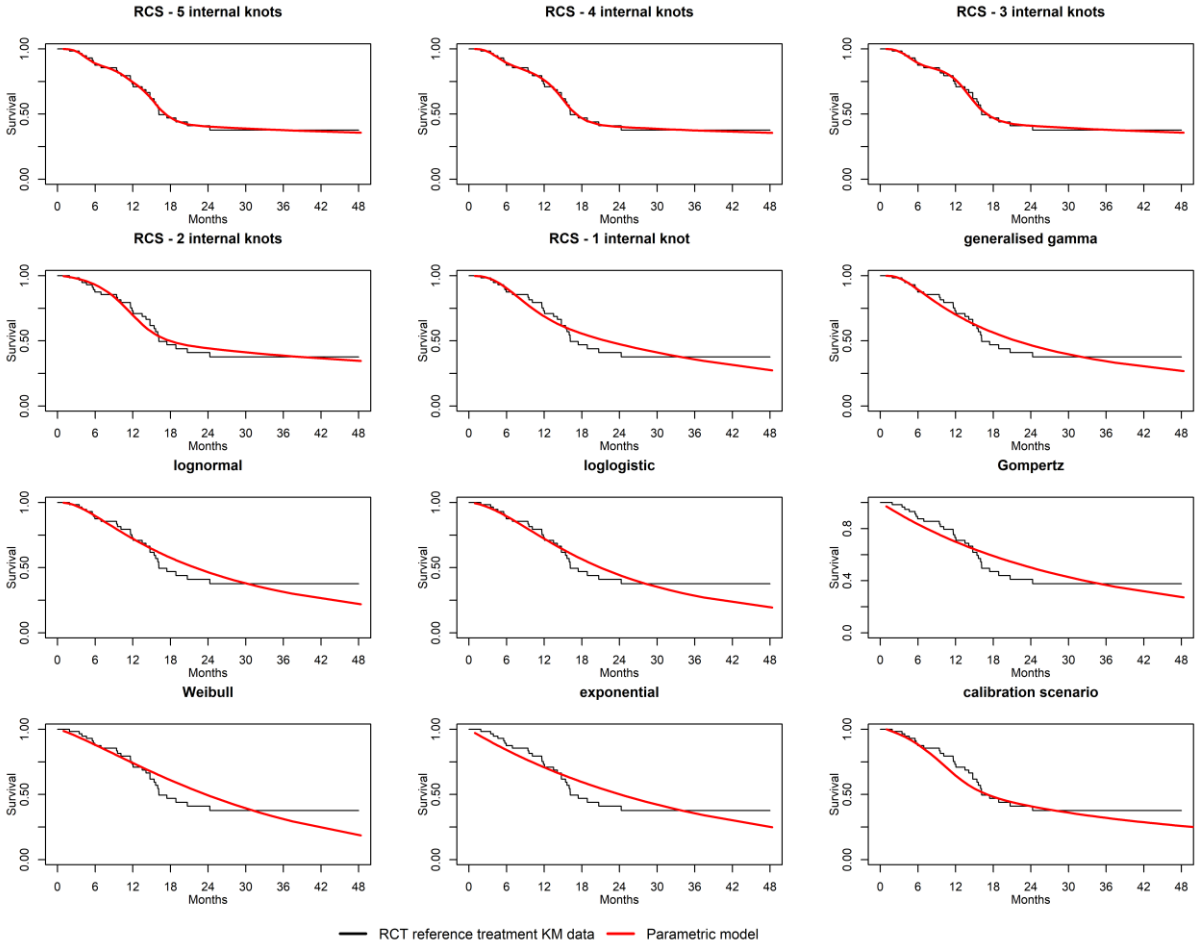


Figure 4.4. Selected models fitted to reference treatment. RCS: restricted cubic spline

The piecewise exponential model with segments from 0-15 and > 15 months was chosen as a compromise to both fit the baseline hazards and relative effects. The piecewise model demonstrated good fit in terms of AIC and BIC (226.2, 229.0), however poorly captured the baseline distribution, substantially underestimating survival throughout both segments (see Chapter 4 Appendix). The number of knots used for the restricted cubic splines (RCS) was informed by signs of overfitting via AIC and BIC values progressively increasing with > 3 internal knots, without improvement in visual fit. Further analyses using RCS were therefore limited to ≤ 3 internal knots. The RCS with 3 internal knots best captured hazards in terms of AIC and BIC values (Table 4.2). Closer visual inspection of

plots, however, showed that the RCS with 3 internal knots displayed some evidence of overfitting the early data due to the closeness with which the volatile early KM estimates were followed (Figure 4.4). The RCS with 2 internal knots did not display this potential overfitting to the same extent, however it did not capture the plateau observed in the tail of the KM estimates as closely. Assuming common hazards between the trial and real-world reference treatment, the calibration scenario‡‡‡ failed to approximate the real-world survival distribution well nor did the RCS with 1 internal knot (Figure 4.4). RCS models with 2 and 3 internal knots were therefore preferred.

Regarding extrapolation, RCS models with 2 and 3 internal knots, which captured the plateau observed in the real-world reference treatment, exceeded background mortality probabilities from approximately 10 to 11 years, assuming linear interpolation between 'steps' seen in general population mortality (Figure 4.5). Therefore, scenarios were incorporated based on the 5-year point used by van der Linden et al. (2015)[49] and my analysis of background mortality, in 2.5-year increments (i.e. at 5, 7.5, and 10 years).

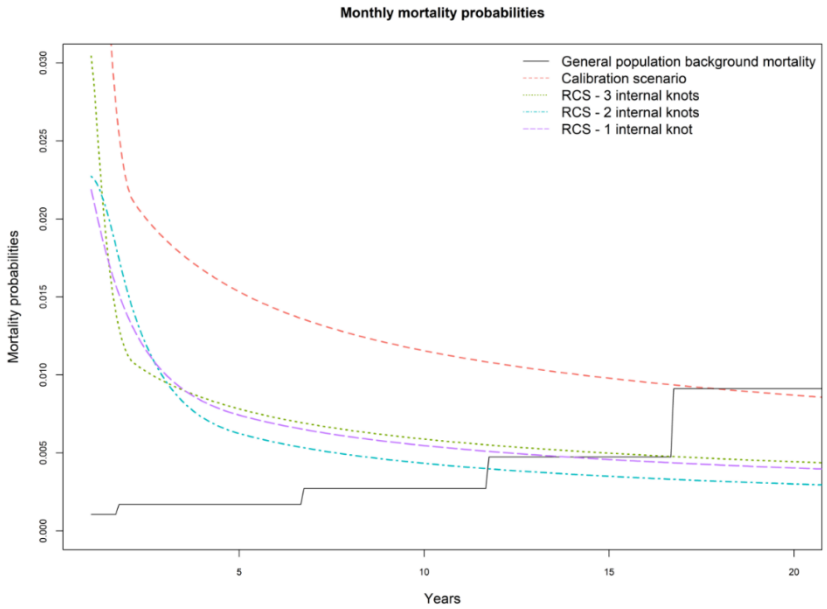


Figure 4.5. Mortality probabilities for background mortality and survival modelling approaches
RCS: restricted cubic spline

‡‡‡ The calibration scenario used the RCS with 3 internal knots fitted independently to each RCT treatment arm based visual inspection and goodness-of-fit (AIC), which was then calibrated to approximate the RWD via a calculated hazard ratio.

Table 4.2. Formal evaluation of model fits

df: degrees of freedom, AIC: Akaike information criteria, BIC: Bayesian information criteria, RW: real-world, RCS: restricted cubic spline, PH: proportional hazards scale, TVC: time-varying covariate(s)

	df	AIC	AIC rank	BIC	BIC rank
RW reference treatment					
RCS 5 internal knots	7	247.1	3	261.9	10
RCS 4 internal knots	6	245.6	2	258.3	6
RCS 3 internal knots	5	244.2	1	254.7	1
RCS 2 internal knots	4	249.8	4	258.2	5
RCS 1 internal knot	3	253.0	8	259.3	8
Exponential	1	256.8	9	258.9	7
Weibull (PH)	2	256.9	10	261.1	9
Gompertz	2	258.6	11	262.9	11
Lognormal	2	250.7	5	254.9	2
Loglogistic	2	251.5	7	255.7	3
Generalised gamma	3	251.4	6	257.8	4
Trial reference treatment					
RCS 3 internal knots	5	1098.9	1	1115.7	2
RCS 2 internal knots	4	1102.5	3	1115.9	3
RCS 1 internal knot	3	1102.0	2	1112.0	1
Trial treatment effects - proportional hazards					
RCS 3 internal knots	6	2379.6	2	2403.9	3
RCS 2 internal knots	5	2377.9	1	2380.6	2
RCS 1 internal knot	4	2398.1	3	2396.8	1
Trial treatment effects - time varying effects					
RCS 3 internal knots, 4 TVC	10	2381.5	6	2422.0	9
RCS 3 internal knots, 3 TVC	9	2379.8	3	2416.2	8
RCS 3 internal knots, 2 TVC	8	2382.7	8	2415.1	7
RCS 3 internal knots, 1 TVC	7	2381.5	5	2409.8	5
RCS 2 internal knots, 3 TVC	8	2378.6	1	2411.0	6
RCS 2 internal knots, 2 TVC	7	2381.2	4	2409.6	4
RCS 2 internal knots, 1 TVC	6	2379.7	2	2404.0	2
RCS 1 internal knot, 2 TVC	6	2384.4	9	2408.7	3
RCS 1 internal knot, 1 TVC	5	2382.5	7	2402.8	1

4.3.2 Relative effects model

Visual inspection of log-log plots (Figure 4.3 top right panel) demonstrated that proportional hazards between arms was questionable, indicating that the impact of time-varying relative effects was warranted. Hazard ratio plots confirmed that relative effects were time-varying (Figure 4.6).

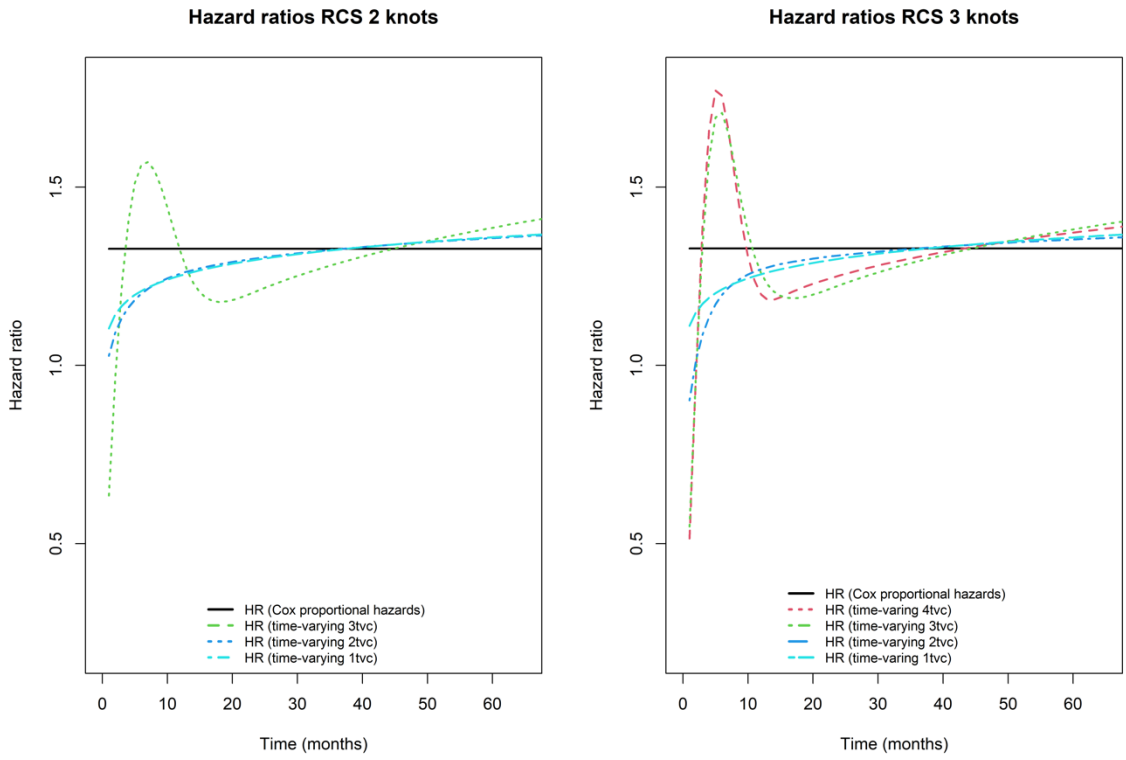


Figure 4.6. Hazard ratio plots using Cox models and restricted cubic splines with varying degrees of freedom
RCS: restricted cubic spline, HR: hazard ratio, tvc: time-varying covariates

Thus, in the case of nonproportional treatment effects, and in the absence of data per comparator to model survival independently, exploration of models to capture time-varying relative effects was required.

While the piecewise model captured distinct relative effects (estimated via a Cox model) within the defined segments, the inability to capture reference treatment models for both the real-world and trial data made it inappropriate for further modelling purposes.

Visual inspection demonstrated that RCS models, as compared with standard parametric or piecewise models, better captured baseline hazards and, consequently, were more appropriate for capturing relative effects. I therefore focussed on RCS models for relative effects. RWD knot locations were used, given the shorter follow-up period, when modelling the trial data to derive appropriate time-varying treatment effects. This enabled placement of treatment effects on the real-world reference treatment model (Figure 4.7).

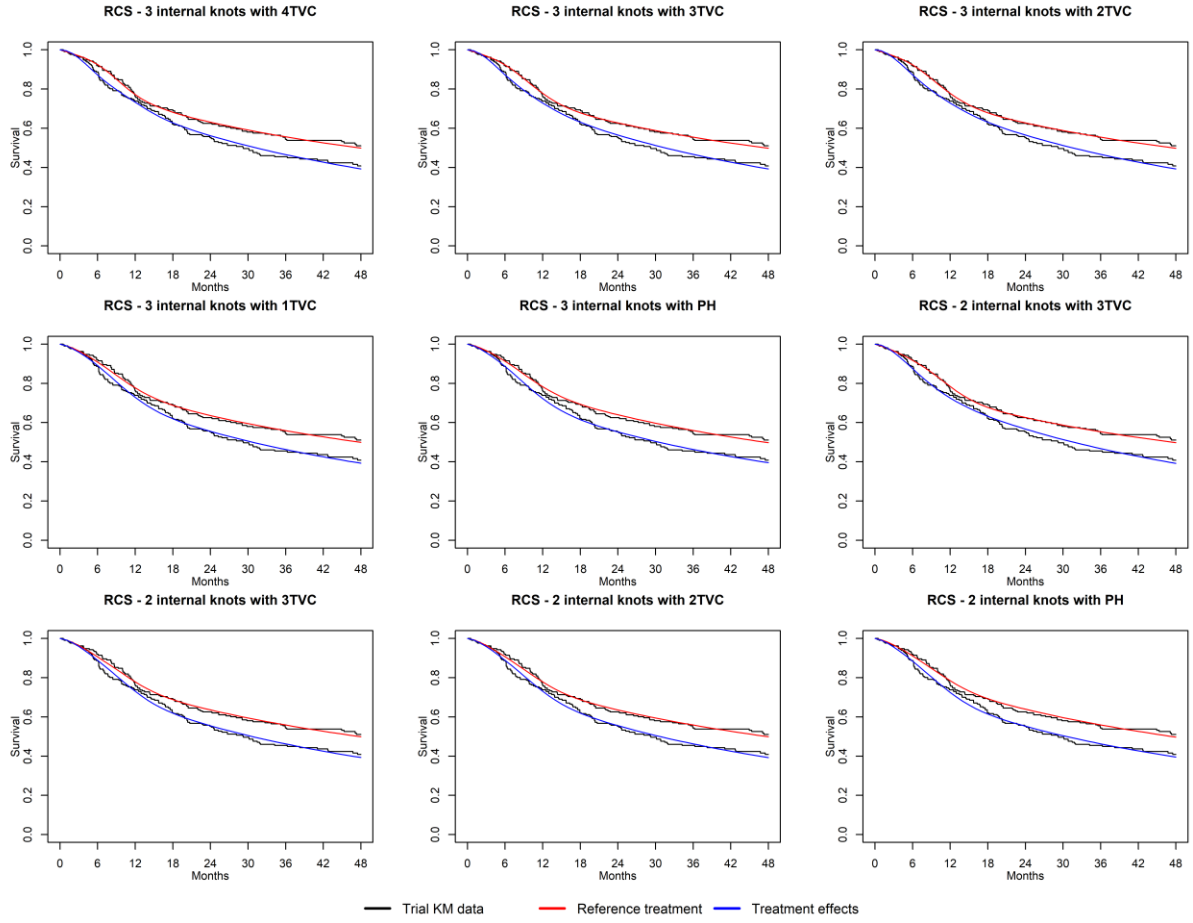


Figure 4.7. Relative effects models using restricted cubic splines

RCS: restricted cubic spline, TVC: time-varying covariate

Visual inspection of models including time-varying relative effects were best captured by the RCS model with 3 internal knots and 4 time-varying covariates, however these were subject to potential signs of overfitting due to the closeness of which the 'non-linear' KM estimates (i.e. nearing/following 12 months) were captured. AIC and BIC values for proportional and

nonproportional effects models were mixed (Table 4.2). By definition, time-varying treatment effects were not captured in scenarios assuming proportional hazards.

Sensitivity to placing treatment effects on higher order terms in different sequences was additionally tested, with estimates found to resemble those derived from the sequential approach as above (Chapter 4 appendix). Given guidance stating that selective knot placement is discouraged[101] and that survival estimates appeared similar, I did not further explore sensitivity to knot placement on cost-effectiveness estimates.

4.3.3 Impact of generalised models on survival estimates and cost-effectiveness estimates

Applying treatment effects to the real-world reference treatment showed different impacts across scenarios. Figure 4.8 shows a selection of resulting curves for extrapolation of overall survival for use in cost-effectiveness modelling with background mortality applied at 5, 7.5, and 10 years. Models with time-varying treatment effects showed greater separation of curves compared with proportional hazards models, reflecting the increased separation of curves observed later in the trial period. Models that did not capture the plateau of the real-world reference treatment (the RCS with 1 internal knot and calibration scenario) estimated decreased survival for both arms, reflective of the hazards observed in the trial as opposed to the RWD.

Cost-effectiveness estimates, in the form of ICERs per QALY, for the various scenarios examined are presented in Table 4.3. RCS model scenarios with 2 vs 3 internal knots resulted in relatively modest differences in ICERs (incremental differences ranged from £18 to £3,691 among scenarios). The assumption of common hazards between the trial and real-world reference treatment via the calibration scenario versus other scenarios which directly modelled the real-world reference treatment data resulted in relatively larger increases in ICERs (incremental differences ranged from £1,819 to £11,429 among scenarios).

Although extrapolation of time varying treatment effects resulted in greater incremental differences in survival between comparators, proportional hazards scenarios resulted in relatively modest increases

in ICERs compared with models incorporating time-varying effects for RCS models with 2 and 3 internal knots (incremental differences ranged from £210 to £2,816 among scenarios).

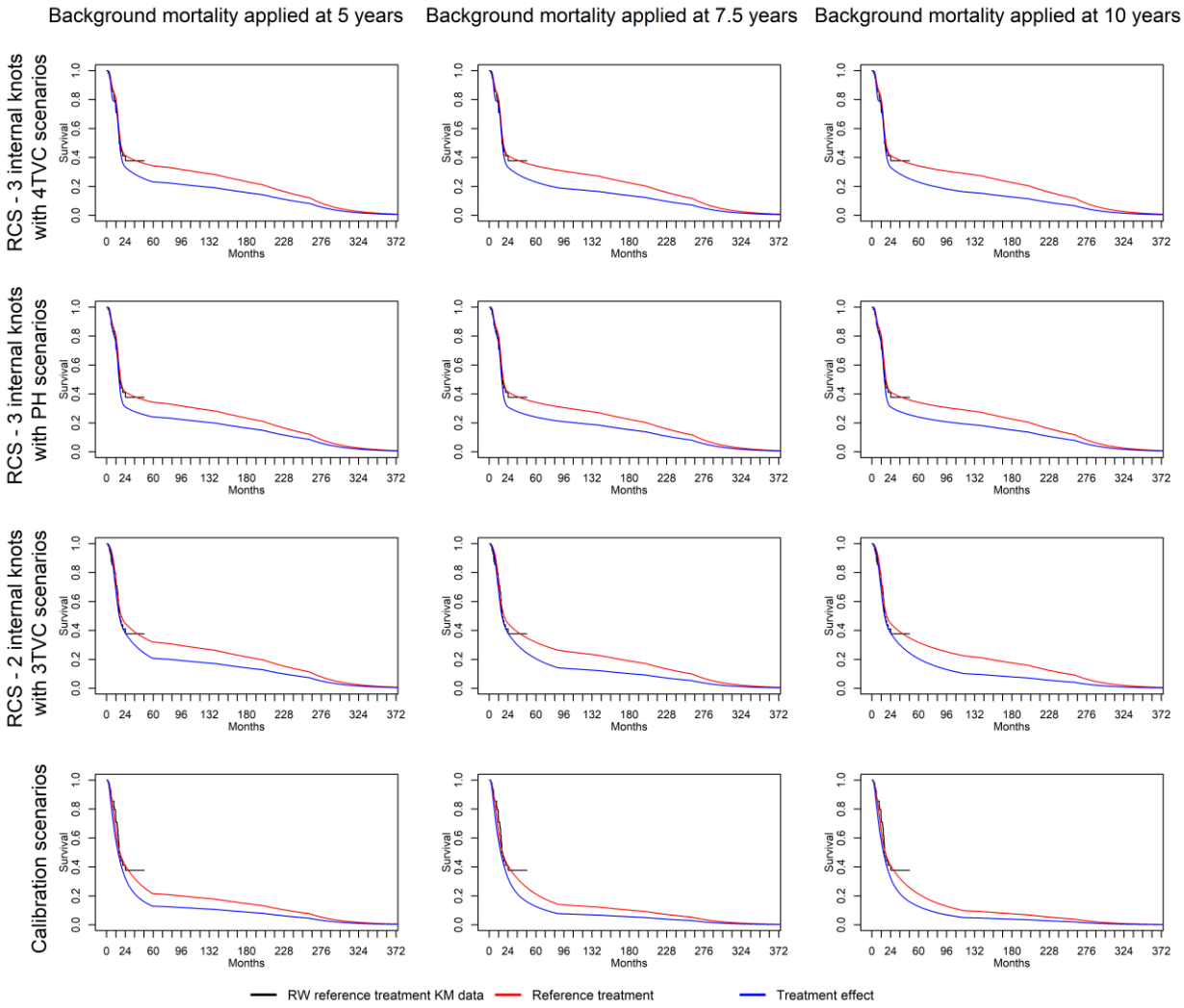


Figure 4.8. Cost-effectiveness modelling scenarios

RCS: restricted cubic spline, TVC: time-varying covariate(s), PH: proportional hazards, RW: real-world

Table 4.3. Cost-effectiveness estimates (ICERs per QALY)

TVC: time-varying covariates, PH: proportional hazards

Background mortality application time point	Restricted cubic spline - 3 internal knots					Restricted cubic spline - 2 internal knots				Calibration scenario
	4 TVC	3 TVC	2 TVC	1 TVC	PH	3 TVC	2 TVC	1 TVC	PH	
5 years	£ 7,104	£ 6,888	£ 7,540	£ 7,405	£ 8,108	£ 6,823	£ 7,476	£ 7,424	£ 8,148	£ 9,967
	(€ 8,261)	(€ 8,009)	(€ 8,767)	(€ 8,611)	(€ 9,428)	(€ 7,934)	(€ 8,693)	(€ 8,632)	(€ 9,474)	(€ 11,589)
	£ 8,642	£ 8,408	£ 9,156	£ 9,025	£ 9,657	£ 8,972	£ 10,479	£ 10,416	£ 11,195	£ 15,813
7.5 years	(€ 10,049)	(€ 9,772)	(€ 10,647)	(€ 10,494)	(€ 11,229)	(€ 10,433)	(€ 12,185)	(€ 12,112)	(€ 13,017)	(€ 18,387)
	£ 10,908	£ 10,781	£ 11,180	£ 11,116	£ 11,390	£ 12,266	£ 14,516	£ 14,463	£ 15,082	£ 22,210
10 years	(€ 12,684)	(€ 12,536)	(€ 13,000)	(€ 12,925)	(€ 13,244)	(€ 14,263)	(€ 16,879)	(€ 16,817)	(€ 17,537)	(€ 25,825)

EUR (€) were converted to GBP (£) using an exchange rate of 0.86

4.4 Discussion

4.4.1 Summary of approach

In this chapter, I presented approaches for generalising trial-derived relative effects to a real-world baseline, in the case of aggregate TTE data with complex hazards, following the below steps:

1. Aggregate TTE data were digitised and recreated estimates were validated against source estimates to ensure subsequent analyses were appropriate.

2. Proportional/common hazards between RCT and real-world reference data as well as RCT treatment effects were assessed. This was done to determine if trial data could be calibrated to approximate real-world baseline data and whether treatment effects could be modelled via a hazard ratio.

3. Given evidence of nonproportionality of treatment effects and nonproportionality between RCT and real-world reference treatment data, common segments between the real-world reference treatment and RCT treatment effects models were used to derive time-varying treatment effects.

4. Time-varying treatment effects were applied to the various segments of the reference real-world distribution.

5. Scenarios examining proportional and nonproportional hazards assumptions for treatment effects and between real-world and RCT reference treatment data were applied in a cost-effectiveness model.

Following these steps provided a structured approach for generalising cost-effectiveness estimates in line with best-practices for survival modelling for HTA, as opposed to potential piecemeal approaches that may exclude on or more critical steps. This is important as omitting one or more critical steps may lead to invalid survival modelling approaches and, consequently, cost-effectiveness estimates.

4.4.2 Summary of key findings

The study showed that, in the motivating example used, capture of real-world reference treatment hazards impacted cost-effectiveness estimates to a greater extent than treatment effect specification. Using calibrated trial hazards, via the calibration scenario, vs direct modelling of the real-world reference treatment resulted in ICERs approximately doubling in some scenarios (e.g. approximately £22,000 vs £11,000 for the 10-year scenario between the calibration and RCS with 3 internal knots). In contrast, assuming proportional vs time-varying relative effects limited increases in ICERs by approximately 25%. Nevertheless, relatively modest differences in cost-effectiveness estimates may well become important for decision-making. Relatively larger differences in cost-effectiveness estimates, as seen among the reference model scenarios in my study, illustrated the importance of appropriately capturing baseline risks for the target population, as has been noted in previous research.[11]

4.4.3 Interpretation of findings

My study demonstrated that handling of complex baseline hazards and non-proportional relative effects, when applicable, can require a number of steps to ensure that estimates are produced in a methodologically valid manner. Naturally, where the proportional hazards assumption holds and there are common hazards between RCT and real-world reference treatments, analyses can be substantially simplified as per approaches outlined in Figure 4.1.

Given the potential number of steps required to generalise trial-derived treatment effects to real-world baselines, I suggest that a balance of model complexity versus impact on cost-effectiveness estimates should be considered, in essence understanding the 'balance of compromises' required to best capture data used to inform health economic evaluations. Understanding this balance can, in some cases, entail a substantial amount of work and potentially result in overly complex models. For instance, regarding the relative effects observed in the RCT from the motivating example, one could argue that assuming proportional treatment effects as a simplifying assumption may be justified on grounds of parsimony. However, without assessing the impact of nonproportional hazards, such a simplifying

assumption can not be sufficiently justified. I thus argue that, regardless of complexity, assessment of common baseline hazards, proportional relative effects, and plausibility of estimates should be explicitly considered to best assess uncertainty and sufficiently inform decision-making.

Although the extended length of the trial used in the example decreased uncertainty in extrapolations, differences between models were non-negligible. The use of immature data, as can be the case at launch of a new technology, can potentially be impacted by approaches used to a much higher degree than what was seen in the current study. Further, as some treatments may demonstrate long-term durable effects, applying common hazards at an assumed timepoint may not be valid. Therefore, it is important to note that the impact of approaches used needs to be evaluated on a case-by-case basis to determine potential impacts on estimates. Although current guidance recommends *"starting from the right place"* in terms of best fitting the observed RCT data when modelling TTE data for health technology appraisal[29], I argue that reliance on RCT data alone for reference treatments can potentially limit external validity of cost-effectiveness estimates. I therefore propose that real-world target population data should, at the very least, be assessed to understand discrepancies between RCT-based and real-world estimates.

4.4.4 Strengths and limitations

Strengths of my study include a systematic approach to survival modelling when applying relative effects to real-world baselines for use in cost-effectiveness analysis. As previous studies in this area identified in Chapter 2 did not report structured steps used to inform chosen survival models [46,48,49], my study provides guidance on selecting appropriate methods. In the motivating example used, I evaluated these methods in the context of complex hazards, thus a variety of approaches were necessarily explored. This makes suggested approaches suitable for a variety of TTE data which may be encountered. Furthermore, the use of aggregate data alone makes the approaches appropriate for use where individual-level data are unavailable.

As various HTA bodies' methods guidance explicitly describe generalisability or external validity as a consideration in decision-making[22–25], the approaches presented can be utilised to formally evaluate

the impact of different baseline risks on cost-effectiveness estimates when faced with complex TTE data. As the use of RWD in HTA increases[109], the presented approaches can be used to supplement analyses conventionally based on risks observed in RCTs alone to better inform decision-making for the specific target population being assessed. This aligns with previous proposals that the formal evaluation of generalisability should be an explicit requirement for cost-effectiveness analysis in HTA.[11] Further, with evolving standards of care, new reference treatments, and/or growing evidence on long-term effectiveness, the flexibility of the approaches allows for iterative evaluation as new information becomes available, as has recently been proposed in the context of 'living' HTA.[110]

There were a number of limitations in my study. First, the relatively large differences in reference treatment survival estimates between the trial and observational study indicated that the generalisability of the trial to the target real-world population may be limited, thus motivating the analysis presented. In practice, where estimates between trial and real-world populations are similar, the approach outlined in this chapter may not be necessary, and generalisability could potentially be inferred by examining survival estimates between sources alone, without the need for a supplemental cost-effectiveness analysis. Nonetheless, the differences in survival estimates between the trial and real-world study used in the motivating example bring the generalisability of the applied trial-based treatment effects into question. Although the use of aggregate data in my study prevented interrogation of treatment effect modifiers to address this, further understanding of potential heterogeneity of effects would be informative for decision-making. It should be noted that the real-world outcomes study from which the real-world reference data were derived had substantial limitations regarding assumed confounding which could not be addressed in my study. Results should therefore be interpreted in terms of exploration of methodological approaches as opposed to any confirmatory findings. Nonetheless, as the use of aggregate data in HTA is commonplace due to data access limitations, my approach remains applicable.

Second, the complex time-varying treatment effects, as represented by the trial KM estimates, may well have been data artefact rather than a representation of true treatment effects. Clinical interpretation would be required to determine appropriateness of modelling time-varying vs

proportional treatment effects. This is important as the increased separation of extrapolations seen with time-varying treatment effects models is to be expected given extrapolation of the later, larger time-varying relative effect. Relatedly, I assumed that KM reference treatment estimates observed in the observational study were valid, however clinical interpretation would be required to assess the clinical plausibility of the observed plateau. Additionally, sensitivity of extrapolations to selective knot locations, informed by clinical opinion or other external evidence, may impact extrapolations and can therefore be considered for decision-making in HTA, as appropriate. However, recently published guidance has noted limitations regarding selective knot placement due to subjectivity, and thus recommends using default knot locations as standard.[101] Further, as my study only evaluated deterministic outputs, inclusion of probabilistic analysis, as would be required for HTA, may have yielded different results if relationships between cost-effectiveness model parameter inputs were non-linear.[111] Nonetheless, given the study was aimed at achieving sufficient fit of parametric survival models to observed values (i.e. approximation of parametric survival estimates to observed estimates), as opposed to confirmation of cost-effectiveness of the intervention, I do not consider this to be a substantial limitation in meeting the study's objectives. Third, my study focused only on the use of flexible spline and piecewise models and I acknowledge that other flexible models (e.g. cure models, landmark models, etc) would have produced different estimates, with impacts on cost-effectiveness estimates unknown. Relatedly, I evaluated sensitivity to placement of treatment covariates on higher-order terms only at a high level to ensure relative consistency of fits, and further evaluation of all possible permutations may have resulted in different estimates. The application of treatment effects to an external data source also brings about issues in terms of covariance used in probabilistic analysis[64], a requirement for HTA.[22,25] Lastly, other extrapolation methods used in cost-effectiveness modelling, such as treatment effect tapering, were not explored.

4.4.5 Conclusions

In conclusion, in this chapter I found that, when attempting to generalise cost-effectiveness estimates by applying trial-derived relative effects to real-world baselines, approaches for capturing reference treatment/baseline survival can be substantially impactful. A balance between model complexity and

adequacy and appropriateness for decision-making should be considered where multiple data sources with complex hazards are being evaluated. The same rigour used for typical curve fitting methods employed to capture trial-based estimates should be applied when aiming to generalise estimates and real-world target population data should be used to inform decision-making.

5 Using real-world survival estimates to inform extrapolation in Health Technology Appraisal

5.1 Background

In Chapter 4, I highlighted the potential impact of survival modelling approaches on cost-effectiveness estimates, and therefore decision-making, using aggregate data alone. In this chapter, I investigate approaches from another perspective, by using individual-level RWD indicatively to inform survival modelling of clinical trial data for use in cost-effectiveness modelling.

Survival estimates extrapolated beyond observed trial data are required for cost-effectiveness analysis used in HTA.[3] Given the relatively short length of follow-up in clinical trials, there can be considerable uncertainty in estimates based on parametric survival models fitted to observed trial data.[26] Extrapolated estimates can play a crucial role in decision-making in HTA, given the lengthy time horizons often required for cost-effectiveness analysis.

Previous research has found that survival models incorporating external registry or background mortality data have improved accuracy of extrapolations up to five years, with long-term accuracy remaining uncertain.[33] While plausibility of extrapolations can be assessed via general population background mortality, this is limited to age and sex-matched survival and could be considered a weak form of evidence for HTA.[35] To better understand uncertainty in extrapolations, individual-level RWD can be used to directly model survival if it can be adjusted to be representative of the HTA target population.[28] While this approach is preferred and allows for various methods for integrating RWD alongside clinical trial data[112], such data may not be readily available as it is dependent on information included in available databases, which can be limited.[85] Where data cannot be fully adjusted to represent a target trial population, NICE DSU TSD 14 states that such data can be used indicatively to understand plausibility of extrapolations, and is preferable to using no external data at all.[28] However, as highlighted in the introductory chapter, there is a lack of clear guidance on how such data should be used to inform plausibility of trial-based extrapolations.[31] Approaches for

applying survival estimates in this context to cost-effectiveness modelling for HTA requires investigation.

In this chapter, I investigate application of real-world overall survival for a specific disease in the context of cost-effectiveness modelling for HTA. I outline a proposed approach for using real-world overall survival indicatively to inform model selection for HTA. I then apply the approach in a motivating example using a previous HTA where survival extrapolations were highly contested among stakeholders to illustrate potential impacts on decision-making.

5.2 Methods

5.2.1 Data sources

I used a CPRD GOLD Primary Care dataset, as outlined in Chapter 3 (June 2020 dataset), to estimate overall survival in individuals with a diagnosis of lung cancer. Briefly, all-cause mortality in individuals with lung cancer was used to estimate overall survival in patients 18 years or older with a recorded diagnosis between 1 January 2000 and 1 January 2019. Additional details are provided in Chapter 3.

For application in cost-effectiveness modelling, I recreated target RCT KM survival data[73] using the algorithm from Guyot et al. (2012)[72] as outlined in Chapter 3. This enabled curve fitting and replication of approaches used in the HTA. Cost-effectiveness model parameters were taken from the HTA submission documents.[34]

5.2.2 Approach to survival modelling

In this chapter, I aimed to provide guidance on using real-world survival data for a specific disease to inform survival modelling of RCT data for HTA. I focussed on application where extensively matching real-world populations to target trial populations may not be feasible due to data limitations. The approach first consisted of selecting appropriate adjustment methods to enable comparison of real-world with target trial survival estimates. If real-world survival data were considered suitable for

providing information on potential long-term survival trends, further analyses were conducted, consisting of:

- right-censoring data at a point corresponding to length of follow-up informed by the target trial, thus creating a right-truncated subset of data resembling follow-up of the target trial
- inspecting data to guide modelling approaches and fitting suitable survival models to the right-censored data
- assessing extrapolations for accuracy against full observed real-world survival to select preferred survival models fitted to trial data

The systematic approach followed is depicted in Figure 5.1.

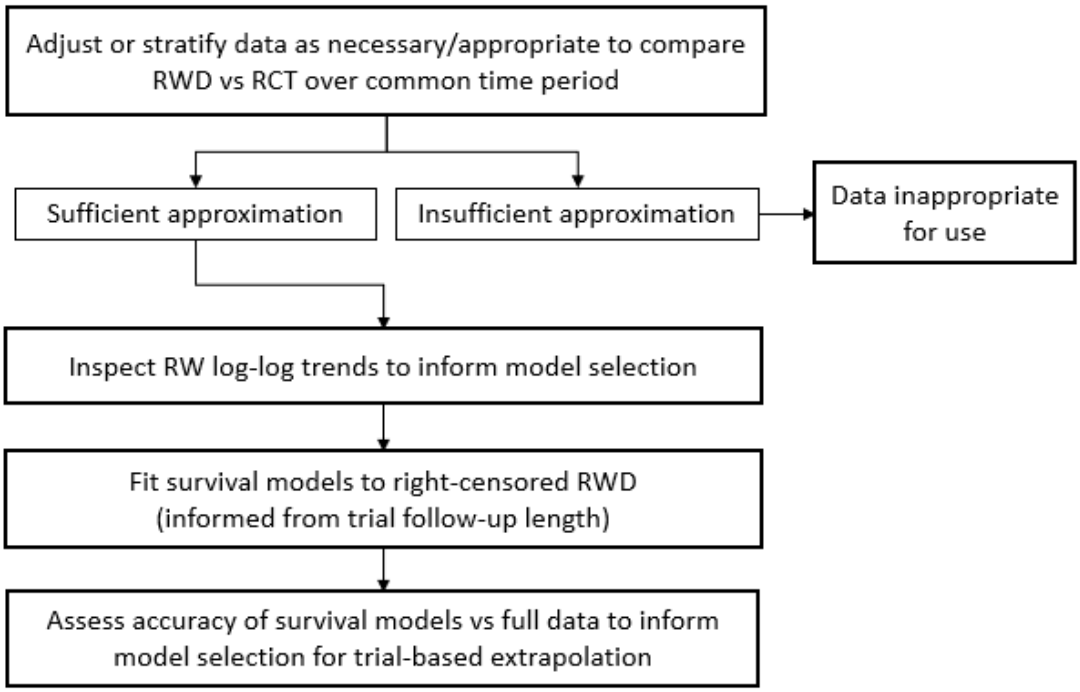


Figure 5.1. Flowchart depicting approach to survival analysis

RWD: real-world data, RCT: randomised controlled trial, RW: real world

5.2.2.1 Adjustment

To best match the real-world population to the target trial population, given available characteristics in the real-world dataset, I first assessed potential adjustment approaches. As described in Chapter 3, the majority of diagnostic codes used to record lung cancer in CPRD GOLD are unspecific and other

characteristics such as staging are not included, thus limiting potential adjustment. I therefore matched the real-world population to the target RCT population using age and sex alone. I visually assessed age and sex strata for proportionality using log-cumulative hazards plots to inform potential approaches (see Chapter 4 Section 4.2.2 for an overview of the use of log-cumulative hazards plots). This was done because violation of the proportional hazards assumption would necessitate stratification to capture different distributions among strata, whereas proportionality would allow for covariate adjustment. This is an important consideration when faced with continuous vs binary characteristics given implications for sample sizes among strata (e.g. fully stratifying by a continuous variable such as age might result in a very small sample size and exclude valuable information). Potential approaches thus consisted of either (1) including characteristics as covariates in fitted parametric models to allow for adjustment if the proportional hazards assumption held, (2) fitting independent models to stratified data if sample sizes were sufficient or if the proportional hazards assumption was violated, or (3) a combination of both adjustment and stratification for various characteristics.

The use of flexible modelling approaches, such as spline models with coefficients placed on higher order terms[91], could have additionally been considered to assess approximation if small sample sizes were a limiting factor for stratification where the proportional hazards assumption was violated. However, the use of this approach would be limiting for subsequent analyses, given the need to fit survival models to unadjusted or stratified right censored data (see Section 5.2.2.3).

5.2.2.2 Approximation of real-world to target trial overall survival

Following necessary stratification and/or adjustment, I assessed approximation to target trial data on the natural scale. Approximation was assessed qualitatively by overlaying adjusted real-world survival estimates with observed target trial survival estimates, given a formal framework for assessing adequate approximation does not, to my knowledge, exist. If estimates were substantially different in terms of median survival or survival trajectories followed very different trends (e.g. indicating time-varying vs constant hazards between real-world and trial sources), real-world

estimates would be deemed unsuitable for informing further analysis. Log-cumulative hazards plots and log-odds plots were used to inform suitability of survival models.[95] As presented in Chapter 4, lognormal and loglogistic distributions are very similar, and suitability of generalised gamma distributions can be assessed via log-cumulative hazards and log-odds plots. Therefore, further variations of log-log plots were not explored.

5.2.2.3 Curve-fitting

Given sufficient approximation as outlined in the previous section, I proceeded to analyse right-censored RWD corresponding to the target RCT length of follow-up at the time of the HTA submission. This was done to understand suitability of survival models for extrapolation vs observed long-term RWD and aid in model selection. If the proportional hazards assumption held among strata, parametric survival models were fitted to the full dataset, given hazards trends would be common and adjustment or stratification would impact absolute estimates rather than survival distributions. If the proportional hazards assumption was violated among strata, it would be necessary to fit independent parametric survival models to capture different distributions among individual strata.

For the real-world dataset, all individuals were retained in the analysis and right-censored if they had not experienced the event of interest by the specified follow-up time based on the RCT. I fitted and examined standard parametric survival models (exponential, Weibull, Gompertz, lognormal, loglogistic, and generalised gamma distributions) and restricted cubic splines with 1 to 5 internal knots (fitted on the log-cumulative hazards scale) for goodness-of-fit, visual inspection, and plausibility. To evaluate accuracy of extrapolations, I considered various metrics. While I assessed formal goodness-of-fit over observed time periods via AIC and BIC values, these metrics only apply to observed data and do not indicate whether extrapolations are accurate.[29] AIC and BIC were therefore not used to assess accuracy of extrapolations. Root mean squared error (rMSE) measures differences in predicted vs observed survival probabilities at multiple time points over the defined time period[113], whereas restricted mean survival time (RMST) measures mean survival over defined time periods.[114] While both metrics can be used to assess accuracy of extrapolations, rMSE avoids

potentially misleading findings in the case of early overestimation and late underestimation, and vice-versa, which can occur if using RMST alone. RMST, however, offers the advantage of being a metric that is easily interpretable and directly applicable to cost-effectiveness analysis, given the use of mean survival estimates in cost-effectiveness models. Given I performed visual inspection to ensure mean survival estimates were not erroneously accurate, I chose to measure accuracy using RMST to make findings directly interpretable for cost-effectiveness analysis, and following approaches used in previous studies[§§§] evaluating accuracy of survival extrapolations.[115–118] I therefore measured accuracy of survival extrapolations vs the full data via differences in predicted vs observed RMST[115], where RMST (in years) was estimated for predicted and observed data for the full time horizon of the observed data. RMST accuracy was defined as the difference between observed (KM estimates) and predicted (parametric models) RMST as per Equation [5.1], with values < 0 representing underestimation and values > 0 overestimation.

$$RMST\ accuracy = \widehat{RMST}_{predicted} - RMST_{observed} \quad [5.1]$$

I further assessed plausibility of real-world survival estimates vs age and sex-matched general population mortality from UK life tables[108] to ensure survival estimates were not implausibly optimistic, that is that extrapolations did not predict improved long-term survival vs the general population.

5.2.3 Cost-effectiveness modelling

I used real-world overall survival in a cost-effectiveness analysis as a motivating example. I chose a previous HTA assessing Nivolumab vs Docetaxel for previously treated locally advanced or metastatic squamous non-small-cell lung cancer where choice of overall survival extrapolation was highly contested.[34] I chose this example as the company and EAG each preferred survival extrapolations with drastically different estimates. This led to the NICE DSU being asked to provide support on selecting the most appropriate approach to guide decision-making in the HTA. It was

[§§§] For this application, RMST is identical to area-under-the-curve if common times are specified.

therefore an ideal example for assessing the potential value of using RWD to supplement RCT data for decision-making.

I developed a simplified cost-effectiveness model based on the HTA, consisting of 'alive' and 'dead' health states (Figure 5.2). I used a partitioned survival modelling approach following the rationale presented in Chapter 4, namely the direct use of modelled survival estimates, a simple but sufficient model structure to capture costs and benefits, and minimal assumptions required to compute health state occupancy. Given I aimed to illustrate impacts on magnitude of changes in economic estimates rather than confirmatory findings of cost-effectiveness, which were predominantly driven by overall survival in the HTA, progression-free survival was not modelled or included in the cost-effectiveness model.

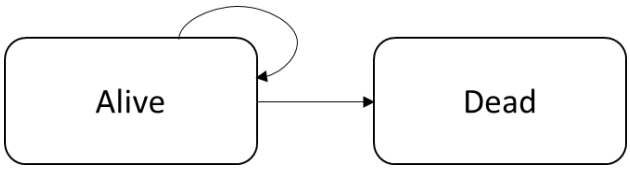


Figure 5.2. Simplified partitioned survival model structure

Where available, I used survival model parameters provided in the submission to create survival distributions[††††]. Where parameters were not presented, I recreated KM estimates using the algorithm by Guyot et al. (2012)[72] as described in Chapter 3. These data were then used to fit survival models and the EAG hybrid KM/exponential approach, hereafter referred to as the piecewise exponential approach.[34] The piecewise exponential approach used KM estimates up to 40 weeks, with an exponential distribution fitted to the observed data beyond 40 weeks and extrapolated across a lifetime time horizon. This approach was chosen by the EAG given distinct hazards trends observed in the RCT data.

[††††] The loglogistic model parameters were presented in the submission as shape = 1.64 and scale = 6.04.

Following the approach used in the submission, and given that the proportional hazards assumption held for overall survival, I applied the hazard ratio‡‡‡‡ from the submission to the Docetaxel arm to estimate Nivolumab survival for all survival distributions. This approach was used for all scenarios aside from the piecewise exponential approach, where Nivolumab and Docetaxel arms were modelled independently. As a simplification, aggregate costs were taken from scenarios presented in the submission, with ranges used to cover sensitivity to ranges estimated via the various survival modelling approaches. Right-censored real-world survival estimates were used to assess suitability of models for extrapolating survival observed in the RCT. Age and sex-matched general population background mortality was taken from UK life tables[119] with probabilities appended at the point at which extrapolated mortality probabilities were exceeded by general population mortality, as necessary. Background mortality was applied following the approach described in Chapter 4 where transition probabilities computed from life tables were utilised. These transition probabilities were applied when modelled survival transition probabilities exceeded those of age and sex-matched general population survival.

I identified preferred models identified from the right-censored RWD to inform extrapolations using RCT data alone. I included scenarios in the cost-effectiveness model for preferred models based on my analysis using RWD, and approaches favoured by the company, EAG, and DSU in the submission. Additionally, I visually compared estimates against long-term follow-up data available since the submission[120,121] as a supplementary analysis to shed light on veracity of various approaches, acknowledging that this information would be unavailable at the time of the HTA.

The outcome of interest was defined as the ICER per LYG given that evaluation of QALYs and progression-free survival fell outside the scope of the current study. A lifetime time horizon was used, where less than 1% of the modelled cohort were still alive at the end of the time horizon for all included scenarios. No discounting or half-cycle corrections were applied given the simplified model structure.

‡‡‡‡ The hazard ratio (HR) for Nivolumab was estimated via a Cox model as stated in the submission (HR = 0.59).

5.3 Results

5.3.1 Study cohort

The real-world population from CPRD GOLD included a total of 38,219 individuals with lung cancer, who had a mean age of 71.4 years (range 18.0-106.0) and were 45.6% female. The RCT population had a mean age of 63 years and were 76% female.[73]

5.3.2 Survival analysis

5.3.2.1 Fitting models to assess suitability of real-world estimates

Visual inspection of log-cumulative hazards plots for age and sex demonstrated that the proportional hazards assumption held among strata, indicating common hazards (Figure 5.3). Given the sufficient sample size[§§§§], I fitted independent parametric survival models to each sex strata. I adjusted for age independently per sex strata to match the mean age observed in the target RCT, given data loss would be substantial under stratification for this characteristic. I adjusted for age using a proportional hazards spline model based on assessed fit to the observed data via visual inspection and formal evaluation (Figure 5.4 and Table 5.1). Survival estimates per sex strata were then reweighted to match the sex distribution observed in the target RCT.

[§§§§] Determination of sufficient sample size was assessed informally, based on the large total number of individuals in the study cohort with a near even split in sex. Visual inspection showing relatively narrow KM estimate confidence intervals and relatively small size of 'steps' in the tails of the KM estimates demonstrated that fitting parametric survival models seemed appropriate.

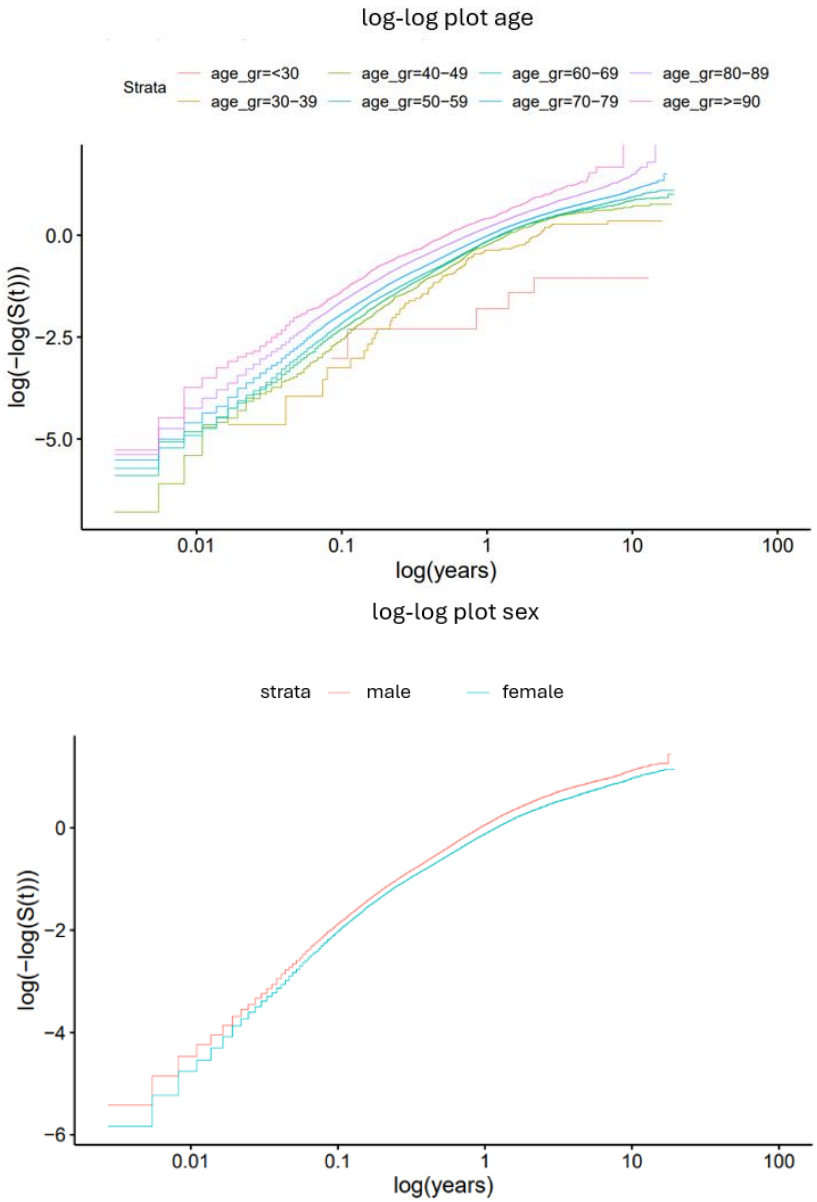


Figure 5.3. Log-cumulative hazards for age and sex strata

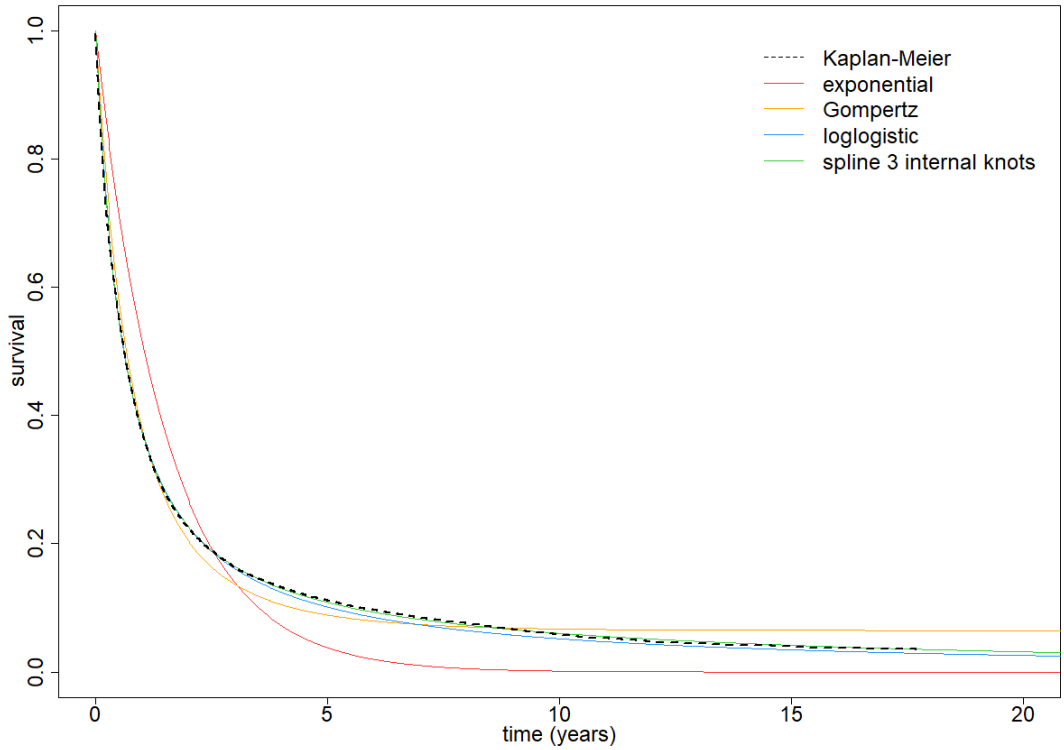


Figure 5.4 Selected survival models fitted to full real-world dataset

Table 5.1. Goodness-of-fit values for full real-world dataset

Survival model	AIC	BIC	AIC rank	BIC rank
exponential	331454.8	331463.3	11	11
Weibull (PH)	321794.1	321811.2	10	10
Gompertz	319329.2	319346.3	9	9
Lognormal	318122.4	318139.5	8	7
Loglogistic	318005.2	318022.3	6	6
Generalised gamma	318117	318142.7	7	8
Spline (1 internal knot)	317970.7	317996.4	5	5
Spline (2 internal knots)	317922.8	317957.0	4	4
Spline (3 internal knots)*	317807.8	317850.6	2	2
Spline (4 internal knot)	317809.6	317861.0	3	3
Spline (5 internal knot)	317783.2	317843.1	1	1

*I used an RCS3 model based on signs of overfitting with the RCS4 increasing in AIC/BIC values with no visual improvement in approximation to the observed data.

The age adjusted, sex reweighted survival estimates matched to the RCT population allowed for visual assessment of RWD vs trial data for plausibility of extrapolations (Figure 5.5). Visual inspection demonstrated that real-world survival estimates did not precisely align with the RCT KM estimates, however sufficiently approximated the RCT data to provide information for understanding survival trajectories beyond the RCT follow-up period. Median survival for the RCT was 9.2 and 6.0 months for Nivolumab and Docetaxel, respectively, compared with 8.5 months for the adjusted real-world lung-cancer cohort. Thus, given real-world median survival fell between the two comparator arms of the RCT and the shapes of the survival distributions were broadly similar, the data were considered appropriate for further consideration.

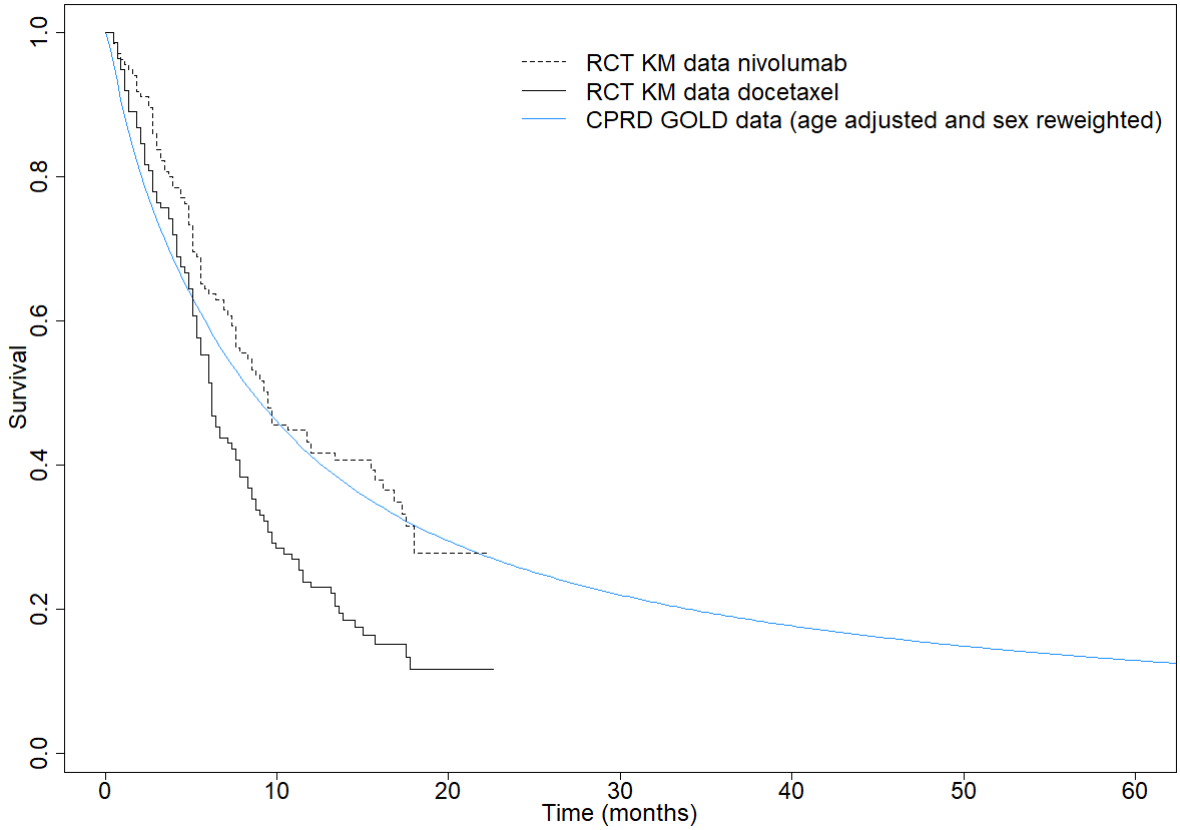


Figure 5.5. Overlaid RCT and age adjusted, sex reweighted real-world survival estimates

RCT: randomised-controlled trial, KM Kaplan-Meier

Age and sex-matched general population background mortality remained lower than real-world estimates across the observed period, however, were similar at the end of the observed period (Figure 5.6). Due to this approximation at the end of the observed period, general population background survival was appended where real-world survival transition probabilities exceeded those of general population background mortality.

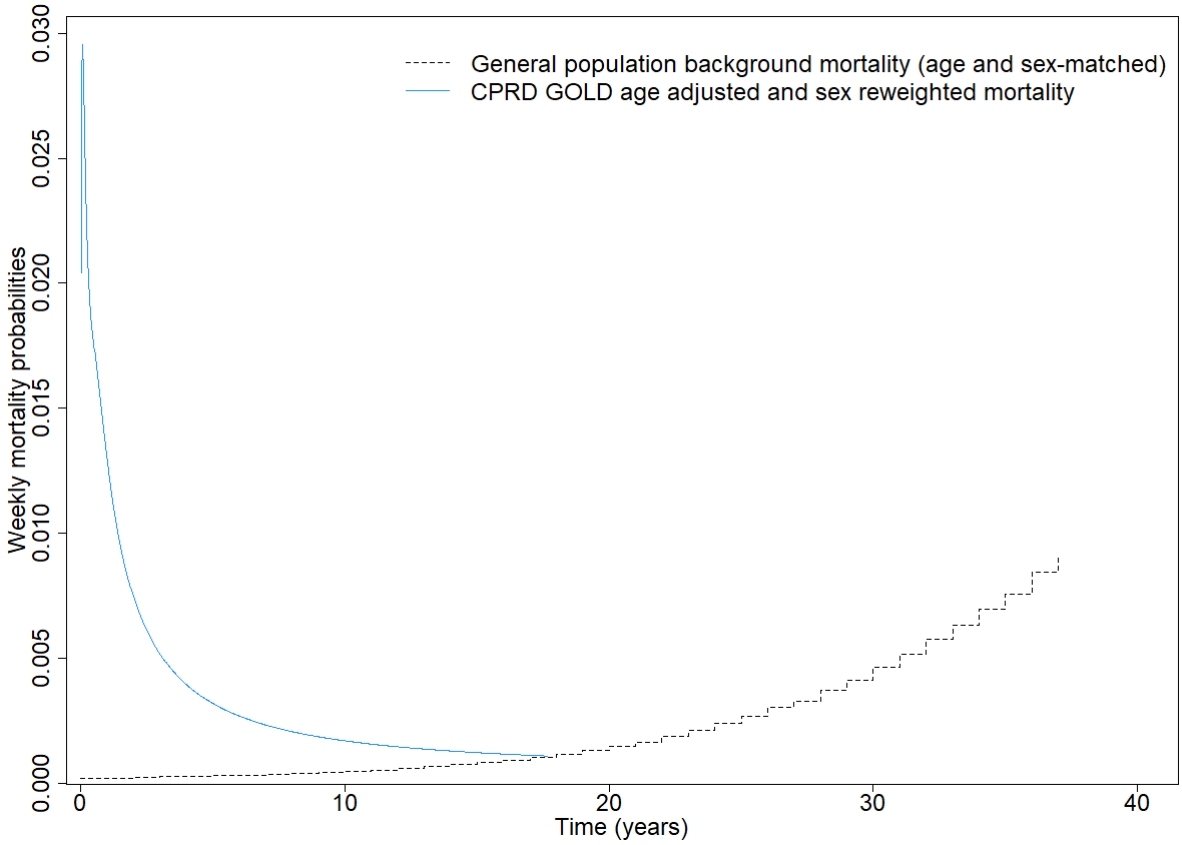


Figure 5.6. Age and sex-matched general population vs age adjusted and sex reweighted CPRD data mortality probabilities

Visual inspection of log-cumulative hazards plots revealed evidence of non-linear trends in the RWD, with the late trend falling beyond the follow-up period of the RCT (Figure 5.7). Log-odds plots revealed that loglogistic (or similar lognormal models[95]) may be appropriate for capturing survival trends (Figure 5.8). This demonstrated that models supporting variable or nonmonotonic hazards may

be required. Notably, the later trend in hazards in adjusted data began between approximately 1.5 to 2 years, falling beyond or near the end of typical follow-up periods of an RCT during time of HTA.

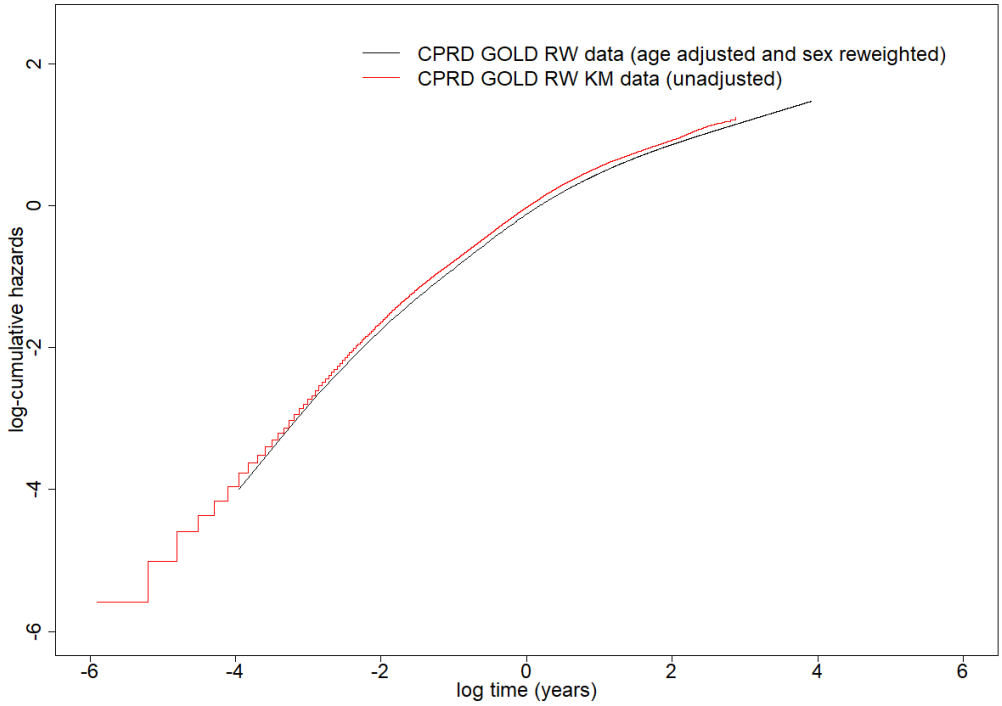


Figure 5.7. Log-cumulative hazards of unadjusted, and age adjusted and sex reweighted RWD
RW: real-world, KM: Kaplan-Meier

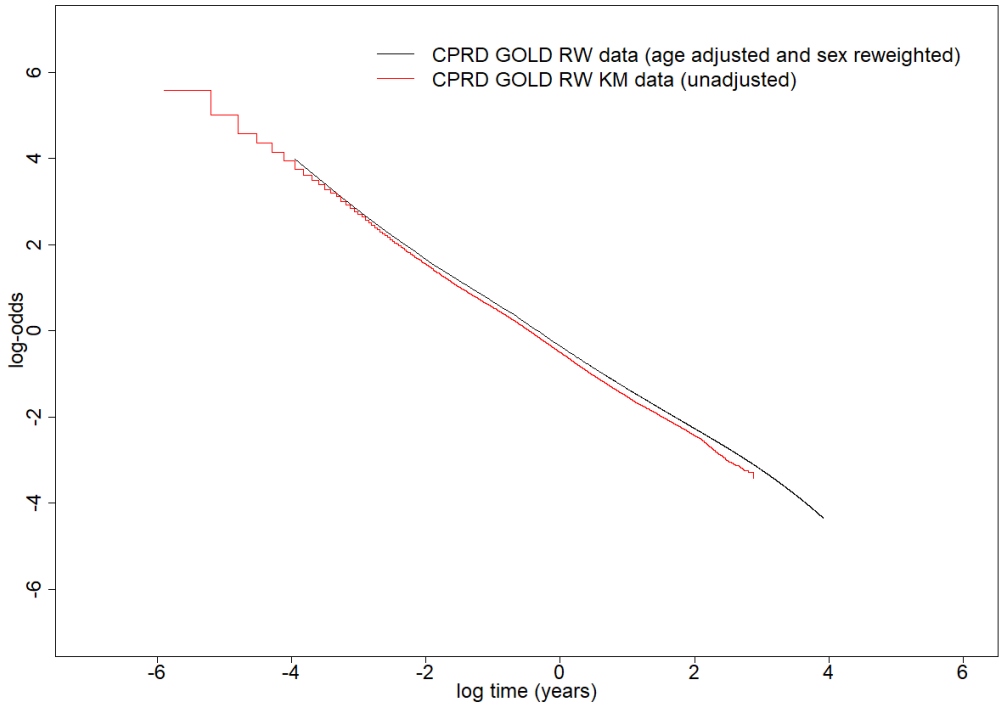


Figure 5.8. Log-odds plots of unadjusted, and age adjusted and sex reweighted RWD
RW: real-world, KM: Kaplan-Meier

5.3.2.2 Fitting models to right-censored RWD

For the real-world survival data, a right-censoring time point of 1.75 years (21 months) was selected, after which the number at risk was zero for both arms of the RCT. I assumed common hazards among strata, given the proportional hazards assumption held upon visual inspection. Had the proportional hazards assumption been violated, it would have been necessary to fit models to individual strata. In such a scenario, if preferred models differed among strata, scenario analyses would be required to understand impacts of models fitted to RCT data for cost-effectiveness modelling. Nonetheless, since the proportional assumption held, I fitted models to the full real-world cohort dataset to inform model selection for the RCT data.

Visual inspection showed that the majority of survival models approximated the observed right-censored data (Figure 5.9). Formal evaluation via AIC and BIC values showed that spline models with four and five internal knots produced the best statistical fit to the observed data (Table 5.2). Visual inspection vs uncensored long-term data, however, showed that spline models underestimated survival, and the loglogistic model performed best (Figure 5.9). This was confirmed via comparison of RMST values where I found that loglogistic, lognormal, and generalised gamma models, respectively, were most accurate (Table 5.3). Best performing models according to RMST as well as the piecewise exponential model preferred by the EAG were then fitted to the RCT data to determine impacts on cost-effectiveness estimates.

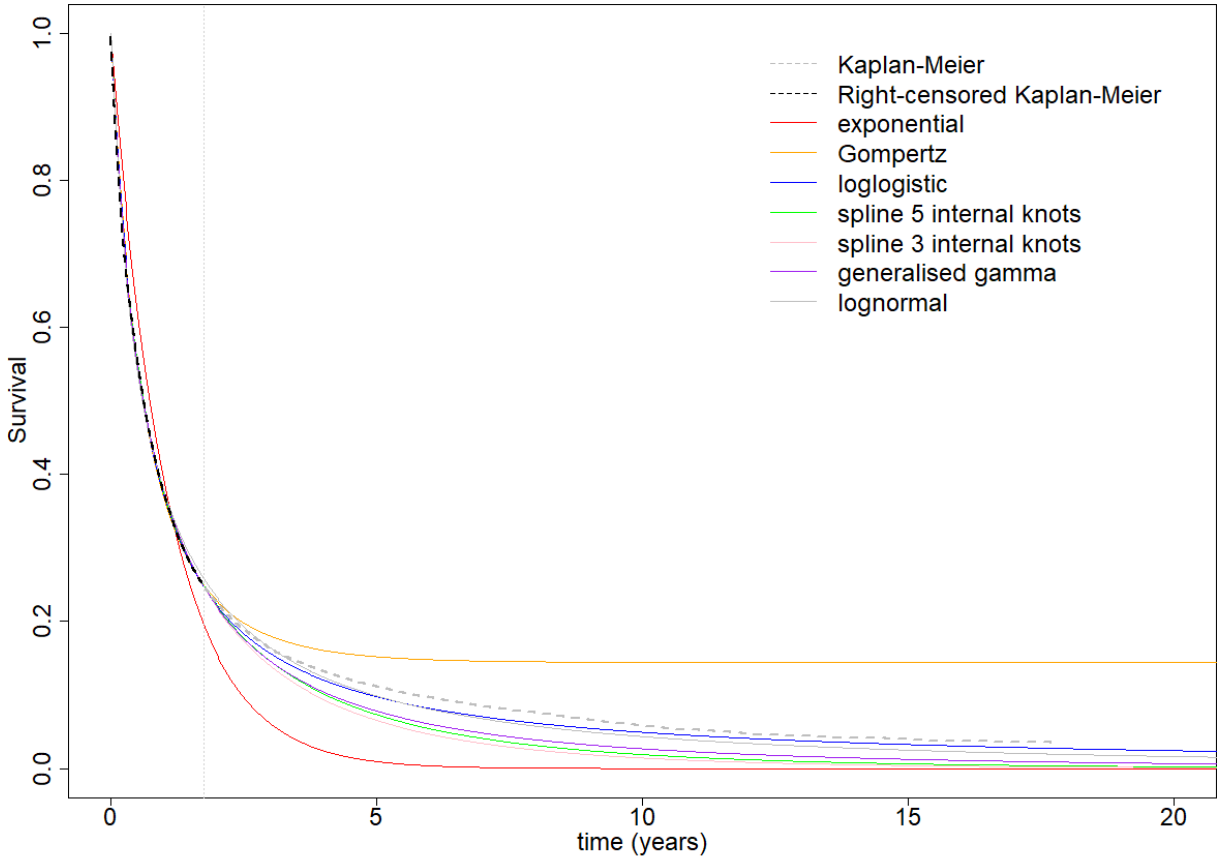


Figure 5.9. Select survival models fitted to right-censored data

Table 5.2. Goodness-of-fit for right-censored real-world dataset via AIC and BIC values

AIC: Akaike information criteria, BIC: Bayesian information criterion, PH: proportional hazards. NB spline models are restricted cubic splines modelled on the log-cumulative hazards scale.

Survival model	**Goodness-of-fit via AIC & BIC**			
	AIC	**BIC**	**AIC rank**	**BIC rank**
CPRD KM estimates	-	-	-	-
exponential	267671.0	267679.6	13	13
Weibull (PH)	265813.2	265830.3	11	11
Gompertz	264900.9	264918.0	9	9
Lognormal	264829.4	264846.5	8	8
Loglogistic	264782.6	264799.7	7	7
Generalised gamma	264664.9	264690.6	3	3
Spline (1 internal knot)	264690.3	264716	6	6
Spline (2 internal knots)	264679.4	264713.6	5	5
Spline (3 internal knots)	264669.4	264712.2	4	4
Spline (4 internal knots)	264617.6	264668.9	1	1
Spline (5 internal knots)	264617.7	264677.6	2	2

Table 5.3. Accuracy of extrapolations vs full data via RMST values

RMST: restricted mean survival time, PH: proportional hazards. NB spline models are restricted cubic splines modelled on the log-cumulative hazards scale.

Survival model	**Accuracy via RMST**		
	RMST (absolute)	**RMST (difference)**	**RMST accuracy (rank)**
CPRD KM estimates	2.091426	reference	reference
exponential	1.084714	-1.00671	10
Weibull (PH)	1.229549	-0.86188	9
Gompertz	3.253657	1.16223	11
Lognormal	1.895289	-0.19614	2
Loglogistic	1.942668	-0.14876	1
Generalised gamma	1.640704	-0.45072	3
Spline (1 internal knot)	1.452578	-0.63885	7
Spline (2 internal knots)	1.437266	-0.65416	8
Spline (3 internal knots)	1.476301	-0.61513	6
Spline (4 internal knot)	1.536097	-0.55533	5
Spline (5 internal knot)	1.549278	-0.54215	4

5.3.3 Cost Effectiveness modelling

Cost-effectiveness results are presented in Table 5.4 for the loglogistic, lognormal, generalised gamma, and piecewise exponential approaches[*****]. ICER estimates for the lognormal (£62,631-£68,735), generalised gamma (£71,647-£78,629), and piecewise exponential (£108,153-£118,693) vs loglogistic (£42,978-£47,116) approach were approximately 46%, 67%, and 152% higher, respectively. Figure 5.10 displays the survival estimates used in the cost-effectiveness model, illustrating the substantial differences in survival between the loglogistic and piecewise exponential models, with the lognormal and generalised gamma falling in between. These survival estimates graphically represent the differences in life-years gained, and consequently ICER estimates, presented in Table 5.4. Additionally, Figure 5.10 illustrates how the loglogistic extrapolated estimates most closely approximated long-term follow-up data since collected from the trial.

Table 5.4. Cost-effectiveness estimates per scenario

Scenario	LYG	Incremental costs range (£)	ICER per LYG (£)
Loglogistic*	1.478		42,978 to 47,116
Lognormal	1.014	63,509 to 69,698	62,632 to 68,735
Generalised gamma**	0.886		71,647 to 78,629
Piecewise exponential***	0.587		108,153 to 118,693

LYG: life-years gained, ICER: incremental cost-effectiveness ratio, *Company Submission preferred model, ** National Institute for Health and Care Excellence Decision Support Unit preferred model, ***Evidence Assessment Group preferred model. Note LYG presented in the submission were 1.446 for the loglogistic and 0.600 for the piecewise exponential approach. The slight discrepancies may be due to application of background mortality for the loglogistic scenario and the use of recreated KM estimates for the piecewise exponential scenario.

[*****] The loglogistic, generalised gamma, and piecewise exponential models were preferred by the company, DSU, and EAG, respectively.

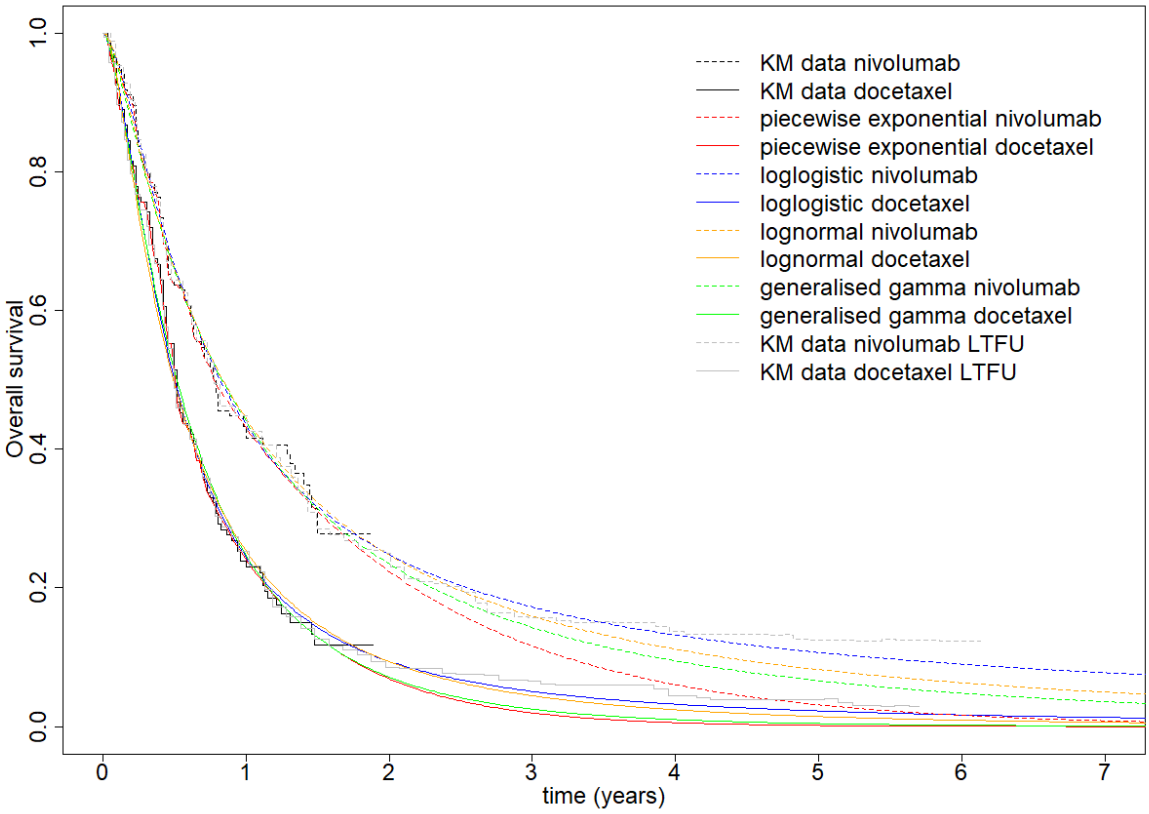


Figure 5.10. Survival models used in cost-effectiveness model
KM: Kaplan-Meier, LTFU: long-term follow-up

5.4 Discussion

5.4.1 Summary of approach

In this study, I assessed long-term real-world survival using RWD from a primary care database in the UK to inform survival model selection for extrapolating observed RCT estimates. Given the real-world dataset lacked characteristics required to adequately match the real-world cohort to the target-trial population, it could not directly be incorporated in cost-effectiveness modelling. I therefore aimed to make the best use of the available RWD by using it indicatively to support selection of appropriate survival models for extrapolating the target trial survival estimates.

A previous HTA assessing treatments for non-small-cell lung cancer[34] was used as a motivating example to illustrate approaches. Acknowledging that long-term follow-up RCT data would not exist at the time of an HTA, I focussed on the use of survival trajectories informed by external RWD to

inform model selection. I fitted survival models to right-censored RWD and assessed accuracy of extrapolations via long-term observed real-world estimates to inform suitability of models for capturing long-term survival trends. I then applied preferred survival models, fitted to the RCT data, in a cost-effectiveness model to illustrate impacts on cost-effectiveness estimates.

5.4.2 Summary of key findings

Log-cumulative hazards plots revealed hazards trends that changed over time, beyond the length of a typical RCT follow-up period, indicating that models supporting variable and potentially non-monotonic hazards may be required.[28] This was further confirmed via inspection of log-odds plots. I found that most survival models closely approximated the observed right-censored RWD on visual inspection. While spline-based models provided the best statistical fit to the right-censored RWD, loglogistic, lognormal, and generalised gamma models were most accurate when compared to the long-term observed real-world survival data in terms of RMST differences. This demonstrated that model selection based on formal goodness-of-fit and visual inspection alone can be limited[30], given the importance of extrapolation. Comparing RCT-based survival estimates to post-HTA long-term follow-up trial data, I found that the preferred survival models selected via the RWD analysis indeed provided the closest survival approximations to the observed data.

In the illustrative cost-effectiveness model, using the preferred loglogistic scenario as a reference, I found that the lognormal and generalised gamma scenarios resulted in 46% (£42,978 vs £62,632) and 67% (£42,978 vs £71,647) higher ICERS[†††††], respectively. While this variation is considerable, the difference when compared with the piecewise exponential scenario was much greater, with a 151% higher ICER (£42,978 vs £108,153)[‡‡‡‡‡].

[†††††] Incremental costs according to constant lower value used. Considering incremental cost ranges, ICERs for the lognormal scenario were 133% to 160% higher vs the loglogistic scenario. Is for the generalised gamma scenario were 152% to 183% higher vs the loglogistic scenario.
[‡‡‡‡‡] Incremental costs according to constant lower value used. Considering incremental cost ranges, ICERs for the piecewise exponential scenario were 130% to 176% higher vs the loglogistic scenario.

5.4.3 Interpretation of findings

Although survival models fitted to RCT data alone are not expected to capture general population background mortality across a lifetime time horizon[35], the piecewise exponential scenario was used in the HTA, in part, on these grounds. This scenario was justified by the EAG because estimates did not exceed general population background mortality during the modelled time horizon, as opposed to other scenarios where extrapolations exceeded general background population mortality in the last decades of life. Although presented, Surveillance, Epidemiology, and End Results (SEER) real-world cancer statistics were not considered in the HTA given uncertainty surrounding applicability of US to UK survival estimates.[34] Due to these perceived limitations surrounding implausibility of long-term extrapolations and questionable generalisability of external data, the EAG based its preferred extrapolation approach on the survival trends observed in the RCT alone. However, given the changing hazards trends observed in the long-term RWD presented in this chapter, the EAG's approach to extrapolating constant hazards from a relatively short follow-up time covered in the RCT alone may have faced additional scrutiny if such evidence had been included in the HTA. Thus, country-specific real-world survival for lung cancer may have been a useful source of external information to better inform plausibility of survival estimates.

A previous study assessing accuracy of extrapolations used in prior HTAs vs long-term follow-up trial data has found that the extrapolated loglogistic model slightly underestimated observed long-term follow-up survival estimates in the HTA used in the motivating example[122], which is consistent with my findings. As this long-term follow-up data would be unavailable at time of HTA, I showed that the use of RWD would be helpful in better understanding plausibility of extrapolations, effectively narrowing down models producing plausible survival extrapolations. In this case, loglogistic and lognormal models which support long tails may have been preferred over constant hazards models if informed by real-world survival. By narrowing down survival models estimating plausible ranges, other important areas of uncertainty, such as persistence of treatment effects[28], could potentially be more thoroughly considered.

In the context of the HTA used in the motivating example, the NICE DSU recommended the generalised gamma model for a number of reasons.[34] First, the DSU noted that the piecewise exponential approach proposed by the EAG did not use all of the observed data for extrapolation, and that using the full observed data may better inform how hazards change over time and be useful for understanding trajectories. Second, the DSU noted that expert clinical opinion was of the view that constant hazards over time, as modelled via the piecewise exponential, was implausible. Third, the DSU aimed to identify an intermediary approach where survival estimates were 'in between' the log-logistic and piecewise exponential model, with the constraint that survival should always be below that of general population background mortality and be greater than progression-free survival. As the loglogistic model exceeded general population background survival near the end of the modelled time horizon, it was considered to be potentially overly optimistic. In-light of this rationale provided by the DSU, my study offers various insights. First, the use of real-world survival may have been useful for informing long-term hazards trends beyond the length of the follow-up included in the RCT and used to identify models capable of capturing these trends. Second, the use of long-term real-world survival may have shed light on the potentially unrealistic requirement for survival extrapolations to be lower than general population background mortality across the entire time horizon, given that the almost 20 years of real-world follow-up did not capture increasing hazards expected in the last decades of life. Third, as the piecewise exponential approach may have been deemed inappropriate if long-term trends observed in real-world survival were accepted, the DSU preferred model may have been different given the rationale of preferring an approach producing estimates 'in between' plausible scenarios. My study demonstrates that, while various survival models can appear entirely appropriate based on observed data alone, the use of long-term external RWD in the example clearly favoured a single model (which strongly aligned with long-term follow-up RCT data collected since the submission). Given the considerable differences in ICERs between approaches, the study exemplifies how sensitive cost-effectiveness estimates can be to extrapolation and emphasises the importance of efforts to inform extrapolations, given the potential consequences for decision-making.

5.4.4 Study strengths and limitations

Strengths of the study include a systematic approach to model selection using a large (N = 38,219) country-specific database. Survival trends based on RWD in my study align with previous research investigating overall survival in lung cancer using CPRD GOLD.[85] This ensured a degree of validity given survival trajectories may differ by country. A follow-up period of 20 years ensured sufficient follow-up to capture long-term survival. My approach differed from more common methods using fully matched RWD cohorts to directly model survival for target trial populations in that RWD were used indicatively to inform survival model selection. This is important given that, in practice, access to RWD can often be restricted and analysts must make do with datasets available to them[30], which may or may not include sufficient characteristics for extensive adjustment. My study thus illustrated an approach for using such data indicatively where other data to inform plausibility of extrapolations are unavailable. Additionally, in contrast to the use of aggregate survival data or point survival estimates, the use of individual-level data in my study allowed for assessment of hazards trends and proportional hazards assumptions to inform adjustment approaches for age and sex.

The study had a number of limitations. First, although individual-level data were available, the study was limited by the information available in the primary care dataset used. Primary care data unlinked to a cancer registry could be subject to misclassification and delayed recording of diagnoses potentially impacting survival estimates. However, as outlined in Chapter 3, previous research has shown minimal differences in lung cancer mortality rates using CPRD GOLD alone vs linked data.[84] Further, as the OMOP CDM was used for analysis, there is potential for information loss when source data were mapped, which could bias results. However, previous research investigating CPRD transformation to the OMOP CDM has shown minimal data loss[123], which provides some assurance regarding validity of estimates from a data quality perspective.

Second, as described in Chapter 3, stratification by lung cancer subtype was not possible in my study given the majority of recorded diagnostic codes for lung cancer in CPRD GOLD are unspecific. Further, lung-cancer survival data used in my study did not account for characteristics such as lung

cancer subtype, cancer staging or Eastern Cooperative Oncology Group (ECOG) Performance Status, among others, given these data are not collected in primary care. Thus, uncertainty regarding representativeness of the real-world population to the target RCT population remains. However, as previous research has shown that the majority of individuals with lung cancer in the UK are diagnosed with non-small cell (80-85%)[87] stage 3 or 4 lung cancer (64-71%)[88], the average individual included in my study is expected to be largely representative. Further, given inspection of previous research suggests similar long-term survival trends among various subpopulations[85], I assumed that model selection based on this relatively homogenous survival trend seen in lung cancer is largely valid. This rationale is reflected in the DSU report comment that different hazard progressions across lung cancer indications may be clinically implausible.[34] Thus, since I focused predominantly on hazards trends rather than absolute estimates in my study, I considered the use of CPRD GOLD to be acceptable in the absence of a dataset including more detailed characteristics. Nonetheless, I acknowledge that the use of survival estimates derived from a real-world cohort fully matched to represent the target trial population would be preferred in place of the approach presented in this chapter.

Third, although CPRD GOLD has previously been shown to be broadly representative of the general population[74], the database has included decreasing numbers of GP practices in England in recent years[124] and uncertainty exists regarding the potential impact this may have had on real-world survival estimates in my study. Nonetheless, similar survival estimates between CPRD Aurum, which includes only GP practices from England, and GOLD in a previous study[86] demonstrates that this uncertainty would have a limited impact on findings. Fourth, although my study benefited from a long follow-up period to derive survival estimates, time periods should be further considered if recent changes in standard-of-care or diagnostics, for example, may have impacted survival expectancy. Fifth, while I assessed both standard and flexible survival models in my study, I fitted splines only on the hazards scale. Given loglogistic and lognormal models performed well, splines on additional scales may have additionally performed well and produced different cost-effectiveness estimates. Nonetheless, given the loglogistic model fit the long term data well, my study effectively demonstrated the potential impact that the use of RWD to inform extrapolations can have in HTA. Sixth, as the study cohort in

this chapter had a large sample size with considerable follow-up, understanding sensitivity of survival extrapolations to sample size, right-censored time points, and censoring distributions would be useful to better understand RWD requirements for supplementing RCT data. I address this limitation in the following chapter. Seventh, my assessment of 'sufficient approximation' of real-world vs RCT survival was performed qualitatively is necessarily subjective. I therefore acknowledge that additional information such as clinical expert opinion would be useful in an appraisal and that the appropriateness of using real-world survival to inform model selection should be assessed on a case-by-case basis to understand validity. Eighth, the finding of proportional hazards among strata in the real-world cohort enabled me to fit survival models to the full right-censored cohort dataset. If non-proportional hazards among strata had been present, indicating heterogeneity in hazards, selection of preferred models may have become more complex. Suitability and impact of my presented approach to inform RCT survival modelling under nonproportional hazards is uncertain.

Lastly, I used a simplified cost-effectiveness model structure to demonstrate magnitude and direction of changes in ICERs, and did not explicitly model parameters such as progression-free survival, adverse events, and resource use. While incremental LYGs were estimated directly from survival analyses conducted in my study, summary incremental cost inputs were taken from the submission as a range across scenarios presented, making them less precise compared to explicit modelling of all parameters. Nevertheless, given the incremental cost ranges presented in the submission cover the scenarios with highest and lowest ICERs (i.e. loglogistic vs piecewise exponential), I argue that my study provides sufficient information to demonstrate magnitude and direction of changes arising from the different survival modelling approaches.

5.4.5 Conclusions

In conclusion, in this chapter I found that real-world overall survival estimates, where adjustment to fully represent a target trial population was limited, were useful for informing extrapolation of RCT estimates. Although RWD could not directly be incorporated in cost-effectiveness modelling due to insufficient representation to the target population, the structured approach presented provides a

framework for using such data indicatively where alternative data to inform plausibility of extrapolations are unavailable. Considering the impact on cost-effectiveness estimates in the motivating example, my study demonstrated that such data may support decision making in the HTA in this context. Given the noted limitations with using such data, the indictive use of RWD to inform survival extrapolations should be evaluated on a case-by-case basis to determine suitability for informing decision-making.

6 Using real-world data to inform survival extrapolations: impact of censoring, sample size, and follow-up

6.1 Background

In Chapter 5, I applied an approach using right-censored real-world survival data to inform model selection when fitting models to RCT data for application in HTA. This approach may be necessary where real-world cohorts cannot be adequately matched to represent the target trial population, thus making direct integration in cost-effectiveness modelling inappropriate. Two recent studies have found that that spline models performed best when assessing predicted vs observed survival estimates using a similar approach to that applied in Chapter 5.[117,118] These studies similarly used right censoring of cancer registry data to identify survival models which performed best in predicting long-term estimates, and provided recommendations on survival models which should be evaluated when extrapolating RCT estimates. These studies, however, noted that the impact of sample sizes and censoring patterns more reflective of RCTs should be explored in future research.[117,118] The appropriateness of transporting findings regarding preferred survival models based on RWD to survival modelling of RCT data is therefore uncertain.

The NICE DSU TSD 21 recommends using sufficiently complex survival models to capture observed data well, however acknowledges that, while splines may capture observed data well, more complex models may not result in appropriate long-term extrapolations.[29] A case in point being splines potentially overfitting particular segments of survival data, particularly when using higher degrees of freedom and smaller sample sizes.[98] Consequently, when the tails of survival data are overfitted, validity of long-term extrapolations may be brought into question. As sample sizes in RCTs can be substantially smaller than those included in large real-world databases/studies, there can be considerably more uncertainty in the tails of KM estimates in RCT data vs RWD. This uncertainty can be seen by the width of confidence intervals and size of ‘steps’ in the tails of KM estimates when comparing RCT vs real-world data. The right-censoring approach applied in Chapter 5, where all individuals who had not yet experienced an event at the introduced follow-up time were censored,

additionally resulted in certainty in the tails of KM estimates not normally seen in RCTs. Loss to follow-up is observed in RWD when individuals transfer out of GP practices, for example[125], however is relatively lower in proportion compared to RCTs, as seen by the discrepancy between numbers of individuals remaining at risk in RCT vs RWD KM estimates over time. In RCTs, censoring is typically distributed over a time frame, and the number of individuals at risk remaining in the last months of follow up is often limited to very low numbers.[29] In Chapter 5, I also used a single follow-up time point for right-censoring, based on length of follow-up in the target RCT, without exploring sensitivity of estimates to this specific time point.

[117,1182998]In this chapter, I investigate impacts of censoring patterns, sample size, and follow-up time when using RWD to inform survival modelling for application in cost-effectiveness analysis. A sampling approach using RWD is used, as opposed to a simulation study, to illustrate a systematic approach for utilising RWD indicatively to inform survival model selection, including additional scenario analyses to that presented in Chapter 5, and to enable comparison with recommendations from previous research. I aim to better reflect uncertainty typically observed in RCTs, building on the previous chapter and the studies by Gray et al. (2021) and Chen et al. (2024)[117,118], when using RWD to inform survival model selection for use in HTA.

6.2 Methods

6.2.1 Data source

As per Chapter 5, I used a CPRD GOLD Primary Care dataset to estimate overall survival, however a more recent dataset (June 2022) was used. Briefly, individuals aged 18 years or older with a diagnosis of lung cancer between 1 January 2000 and 1 January 2019 were included. Additional details regarding the dataset and inclusion criteria are provided in Chapter 3.

Given the aim of this chapter was to make real-world survival data more reflective of typical RCT survival data, I chose to use a subset of individuals based on age. This was based on the assumption that real-world survival estimates would have a greater degree of heterogeneity than typically observed in an RCT.[117] This increased heterogeneity in real-world vs RCT data may occur because

RCT selection criteria can exclude individuals with comorbidities and elderly individuals.[10] In an attempt to reduce heterogeneity, I restricted analysis to a subset of individuals in the real-world dataset. Therefore, the cohort was stratified into 10-year age bands. As RCTs typically enrol younger individuals[126], with a 9.75 year lower median age between trial and population median age found in previous research[127], the 10-year age band one stratum below the largest age band was chosen for further analysis.

Given the aim of the study was to reflect censoring patterns, sample sizes, and follow-up times observed in typical RCTs, scenarios were conducted to address (1) smaller sample sizes observed in RCTs, (2) shorter duration of follow-up observed in RCTs, and (3) distributed rather than uniform right-censoring observed in RCTs.[117]

To reflect smaller sample sizes typically observed in RCTs, individuals from the real-world 60-69 years old stratum were randomly sampled with replacement to produce sample sizes of 150, 300, and 1000. A scenario including the total stratum was additionally included for comparison.

To cover a range of follow-up times often observed in lung-cancer RCTs[101], individuals who, by 12, 24, and 36 months of follow-up, had not yet experienced the event of interest (death) were censored, thus creating three distinct follow-up time scenarios. Absolute times, rather than times expressed as a function of survival percentage, were chosen due to their relevance for HTA. This is because the length of follow-up included in data locks for pivotal RCTs at time of submission is typically used for extrapolation. Gray et al. (2021), for example, used times corresponding to 20%, 35%, and 50% survival for their non-small cell lung cancer (NSCLC) cohort, resulting in absolute times between 3 and 13 months.[117] These durations may be considered relatively short when compared to data locks for pivotal RCTs at the time of HTA submission.

In the first censoring scenario, individuals not yet experiencing an event were uniformly right-censored at the respective artificial follow-up times. In the second scenario, individuals not yet experiencing an event at the respective follow-up time were censored across 50% to 100% of total follow-up at randomly introduced times. In the third scenario, individuals not yet experiencing the

event at the respective follow-up time were censored across the total follow-up time at randomly introduced times excluding the first month, assuming censoring is generally not expected to be high within the first month of an RCT.

For all permutations of above scenarios, cohorts were sampled 1000 times for analysis. For the scenario where the full sample was used, the sample size remained fixed while distributed censoring times were introduced randomly per sample drawn.

6.2.2 Survival analysis

For survival analysis, parametric survival models were fitted to each randomly drawn sample. Standard parametric models consisted of the exponential, Weibull, Gompertz, loglogistic, lognormal, and generalised gamma distributions as per those recommended in NICE TSD 14.[28] Flexible parametric models consisted of restricted cubic splines with 1, 2, and 3 internal knots modelled on the log-cumulative hazards, log-cumulative odds, and inverse normal scales.[91] Splines were limited to 3 internal knots following rationale outlined in Chapter 4 on selecting sufficient number of knots and to limit excessive computational burden. As per current guidance[29,101], default knot locations were used and specified as quantiles of log uncensored event times.[91] A total of 15 parametric models were therefore fitted per randomly drawn sample. Models not converging for individual draws were discarded and excluded as contender models. KM estimates were used to enable comparison between observed and predicted estimates.

Fitted parametric survival models were extrapolated up until the maximal follow-up time observed for each randomly drawn sample. In contrast to Chapter 5, rMSE was chosen to assess accuracy (and determine preferred models) rather than RMST, given that visual inspection across the large number of draws was not feasible. As described in Chapter 5, without visual inspection, early overestimation and late underestimation and vice versa, could result in erroneously accurate RMST estimates used for preferred model selection. Goodness-of-fit was not used to determine preferred models given AIC and BIC values do not provide information on the extrapolated estimates beyond the observed data.

Model estimates were therefore assessed for accuracy using rMSE of survival estimates calculated at monthly intervals using Equation [6.1] to select preferred models per draw as,

$$rMSE = \sqrt{\frac{\Sigma(\hat{S}(t)_i - S(t)_i)^2}{n}}\ , \qquad [6.1]$$

where $\hat{S}(t)$ is the predicted and $S(t)$ the observed survival percentage at *i* monthly intervals, and *n* the total number of monthly timepoints. Thus, the parametric survival model with the lowest rMSE was defined as the preferred model per draw. For each model, the total number of draws in which models were preferred were summed and presented as a proportion of the total number of draws (1000) per scenario.

Given RMST is directly applied in cost-effectiveness modelling (as opposed to statistical measures such as rMSE, AIC, and BIC), RMST was additionally computed to supplement rMSE values to provide insights regarding potential impacts on cost-effectiveness estimates. Per parametric survival model and KM estimates for each draw, RMST was calculated across the total follow-up period (which varied depending on randomly selected individuals per draw). RMST accuracy was defined as the difference between observed (KM) and predicted (parametric models) RMST as per Equation (5.1) in Chapter 5, with values < 0 representing underestimation and values > 0 overestimation. RMST is presented graphically using boxplots to show trends among scenarios including median estimates, interquartile range, and range for the various parametric models fitted.

As supplementary analysis, I compared AIC and rMSE in terms of proportion of preferred models between per draw to provide insights regarding the use of AIC for model selection in the absence of other metrics or approaches.

6.3 Results

6.3.1 Study cohort

A total of 40,850 individuals were included in the full dataset. Mean age across the dataset was 71.36 years. Age band 70-79 included the largest number of individuals (14,971 (37%)), with the 60-69 stratum being second highest at 11,135 (27%). Given the earlier rationale provided in the methods (i.e. that trial cohorts are 9.75 years younger than real-world populations), the 60-69 stratum was used for analysis. This subset had a mean age of 64.98 years (IQR=63-67) and was 45.1% female.

6.3.2 Survival analysis

For the full 60-69 stratum, median survival was 9.44 months (95% CI = 9.11 to 9.76) and maximum follow-up time was 19.45 years (Figure 6.1).

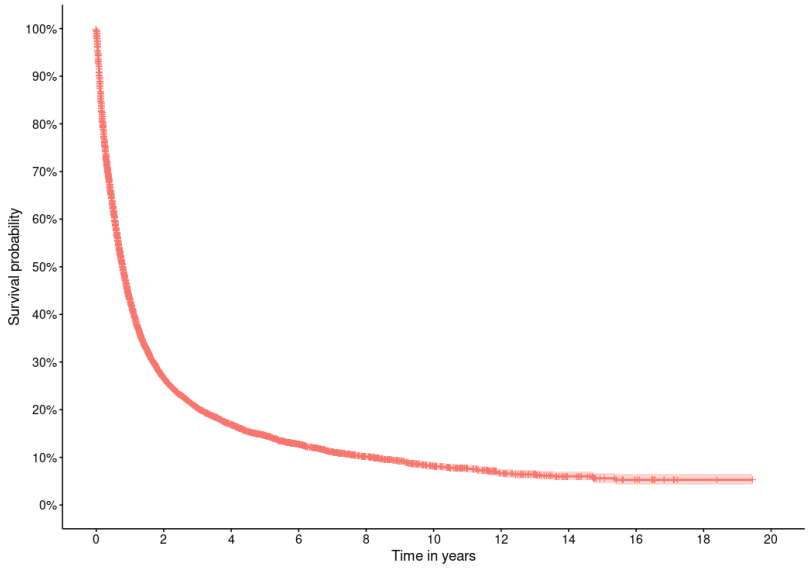


Figure 6.1. Overall survival for the 60-69 stratum

A summary of preferred models, according to the proportion of parametric models with lowest rMSE and AIC across draws, and lowest mean RMST difference across draws is presented in Figure 6.2 for all sample size, follow-up time, and censoring distribution scenarios. The proportion of parametric models with lowest rMSE across draws are presented graphically in Figure 6.3 to Figure 6.6 and numerically in the Chapter 6 Appendix. Spline odds models, lognormal, and loglogistic models broadly performed best according to rMSE across scenarios. Spline models performed best with uniform right-censoring scenarios while lognormal and loglogistic models performed best for

distributed censoring scenarios, with the exception of the Gompertz model performing best for two 36-month follow-up time scenarios (n=1000 and full sample partially-distributed censoring scenarios). Regarding the impact of sample size, smaller sample sizes (n=150 and n=300) displayed more variability, compared with larger sample sizes (n=1000 and full sample), however preferred models across scenarios remained largely consistent. For follow-up time scenarios, preferred models for distributed censoring were generally the loglogistic model for 24- and 36-month scenarios, while the lognormal model was preferred for 12 months scenarios. Otherwise, preferred models remained relatively consistent.

RMST estimates for all draws across scenarios were assessedvia boxplots depicting median, interquartile range, and ranges (see Chapter 6 Appendix). Median RMST estimates followed rMSE estimates in terms of accuracy, however showed substantial overlap in estimates between preferred spline, lognormal, and loglogistic models. Decreased accuracy was seen overall for distributed censoring vs right-censoring approaches and for shorter vs longer follow-up periods. Larger sample sizes decreased variability as seen by RMST ranges, however resulted in decreased accuracy.

Preferred models identified via AIC were inconsistent with preferred models identified via rMSE (see Figure 6.2 and Chapter 6 appendix for graphical depiction of all models).

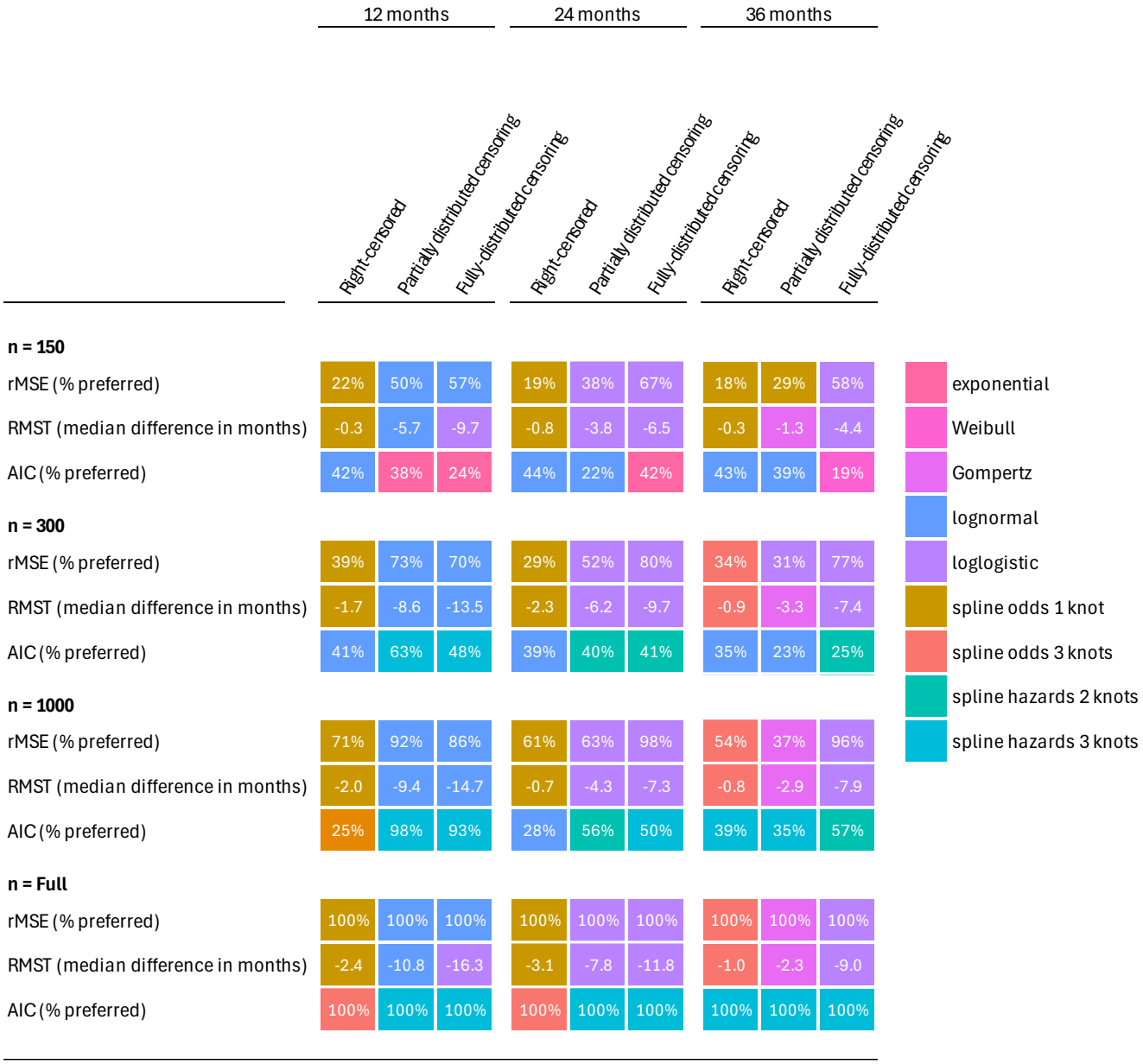


Figure 6.2. Summary of preferred model per scenario for rMSE, RMST, and AIC

rMSE: root mean squared error, RMST: restricted mean survival time, AIC: Akaike information criteria

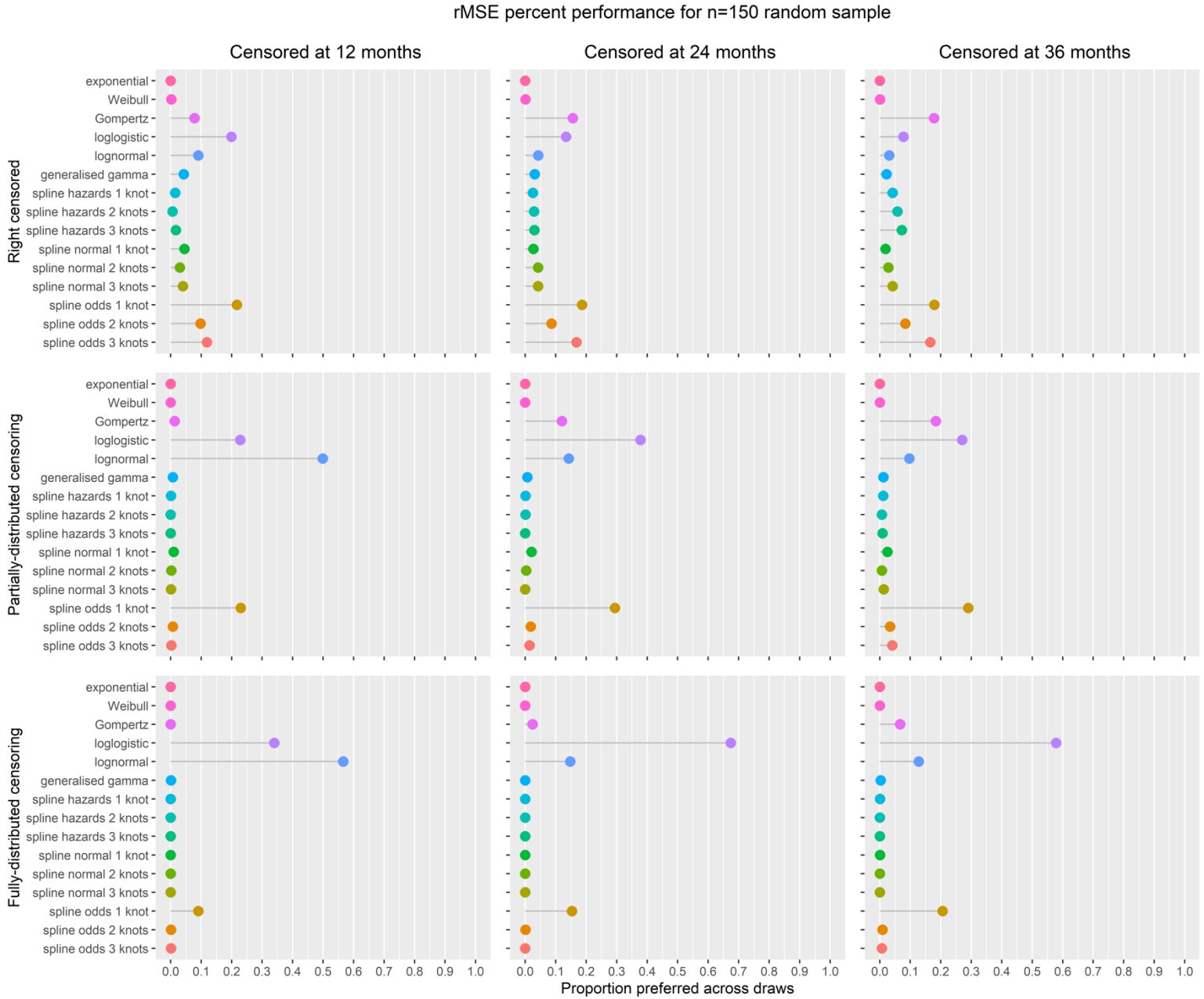


Figure 6.3. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=150 sample size, for each scenario

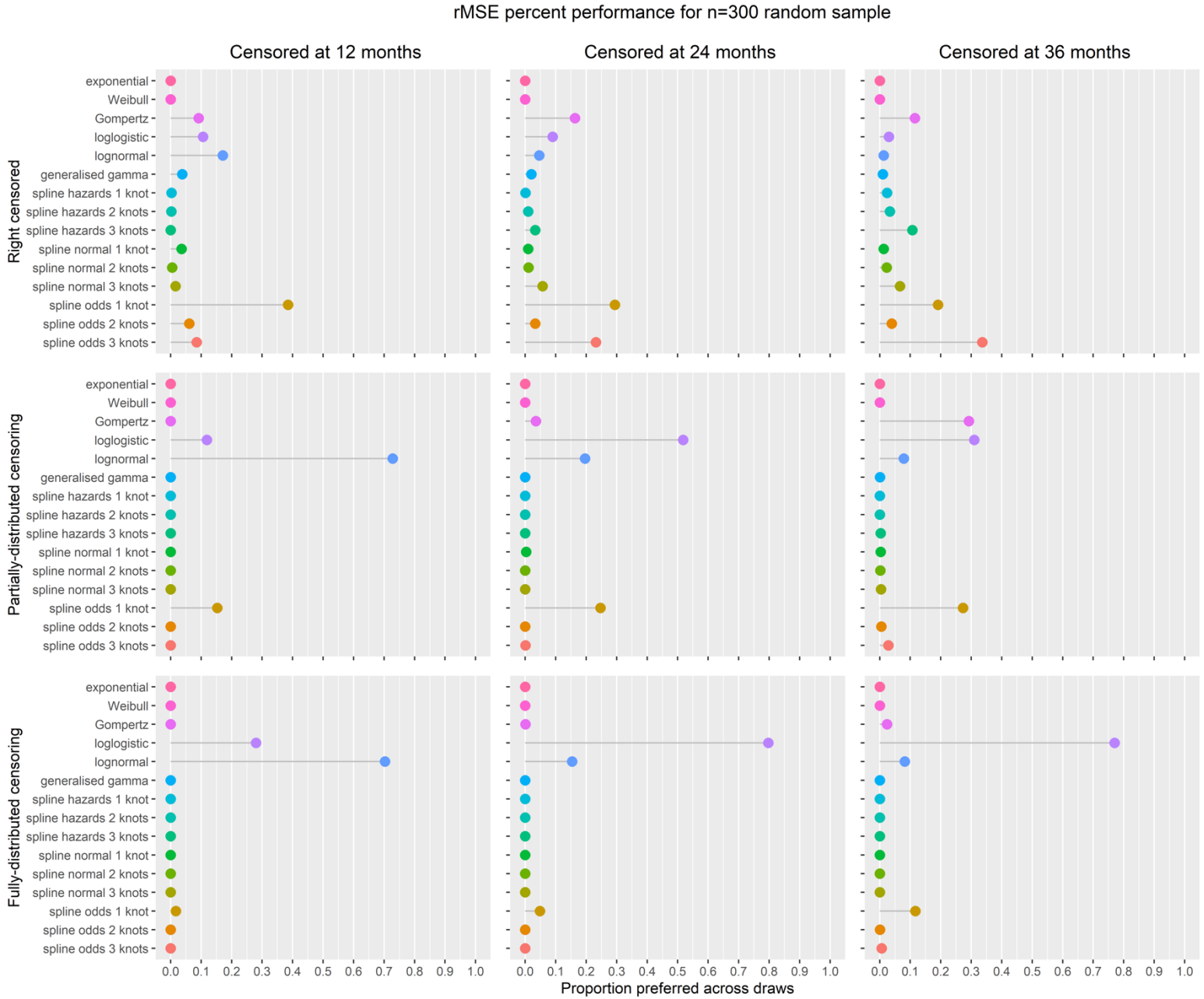


Figure 6.4. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=300 sample size, for each scenario

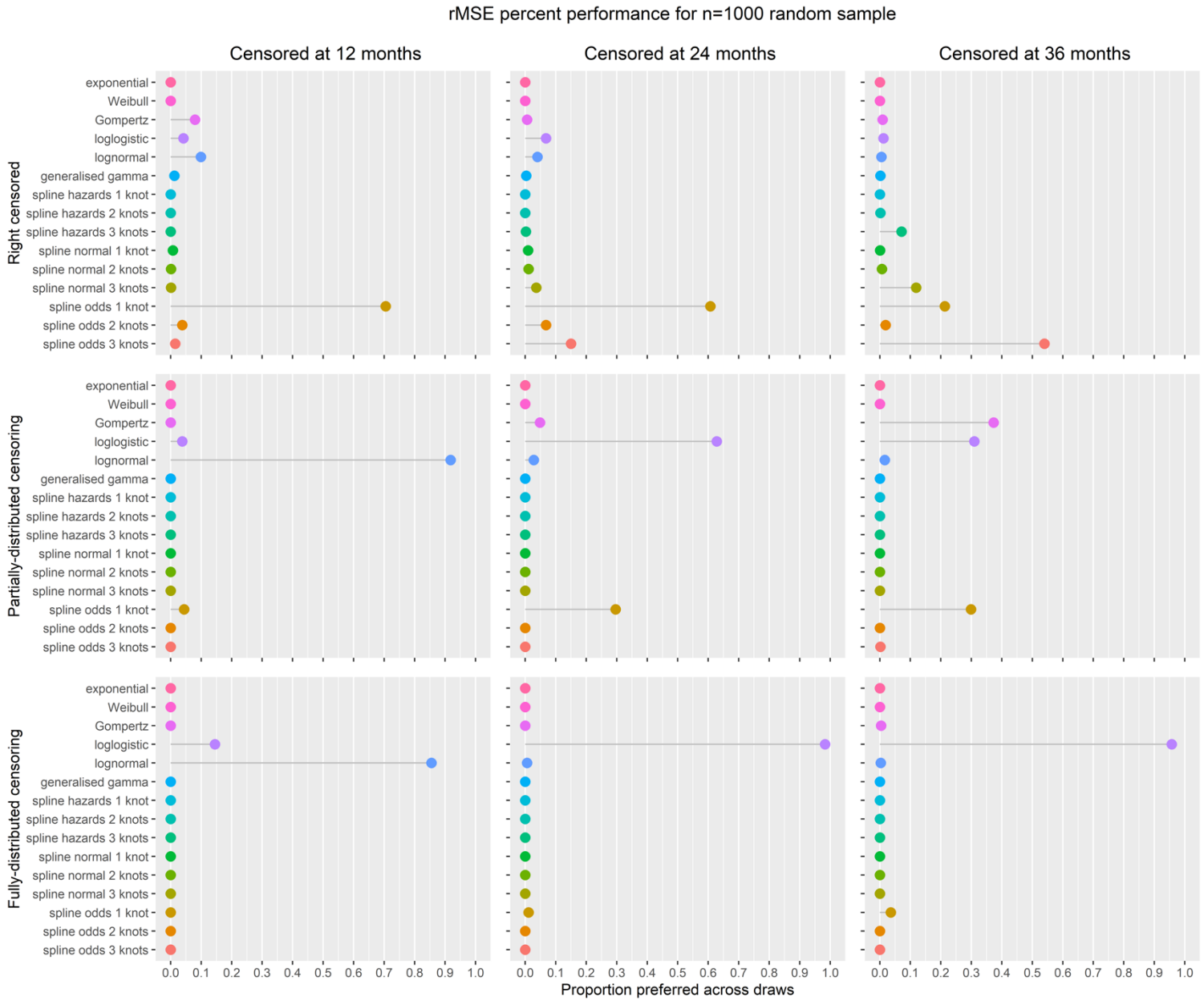


Figure 6.5. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=1000 sample size, for each scenario

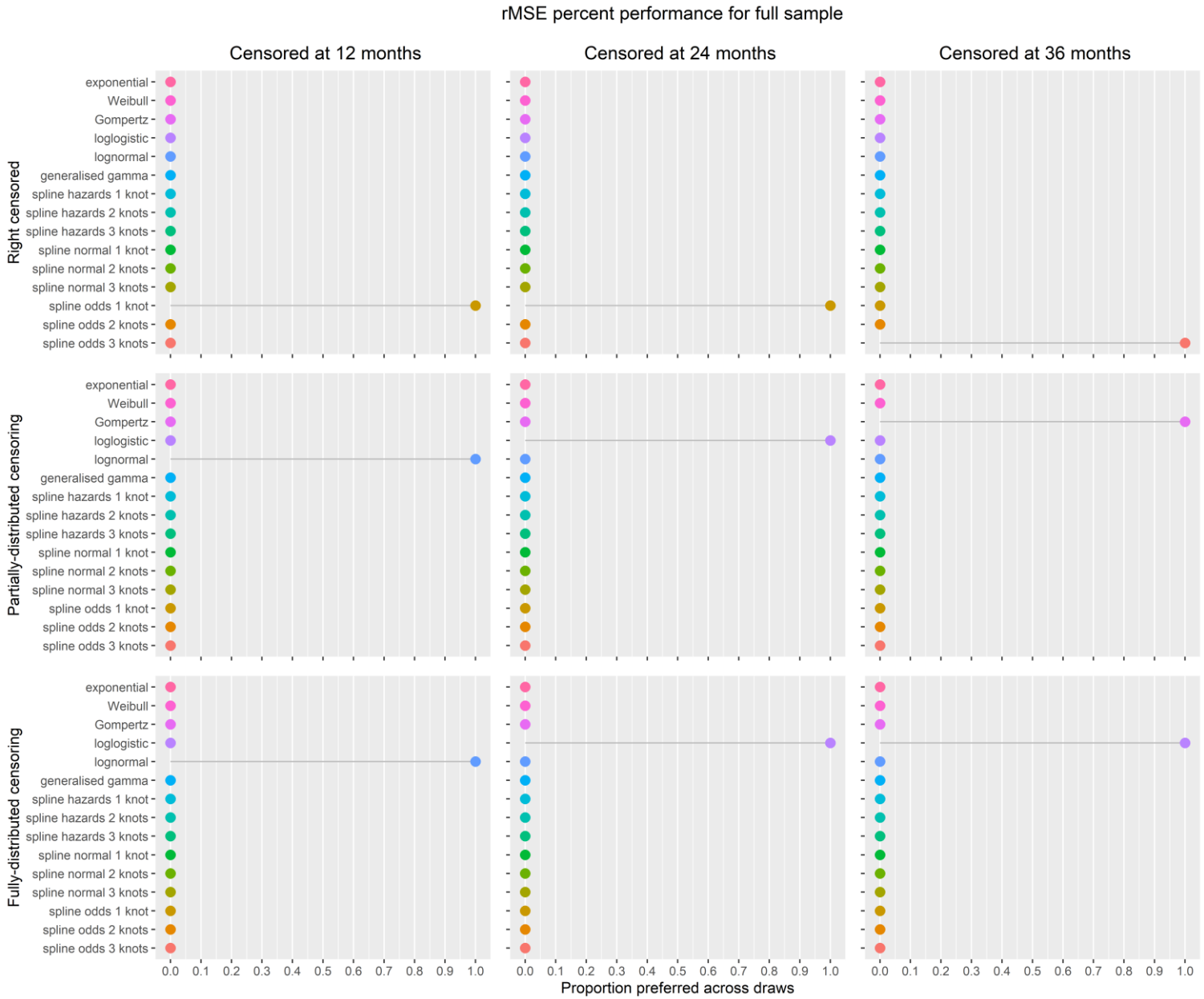


Figure 6.6. The percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the full sample size, for each scenario

6.4 Discussion

6.4.1 Summary of approach

In this chapter, I investigated the impact of censoring distributions, sample size, and follow-up time on survival extrapolations when using RWD to inform model selection. I used a lung cancer cohort using a primary care dataset in the UK to estimate overall survival. Given RWD is different from RCT data in terms of sample size, follow-up, and censoring distributions, I expanded on the approach used in the previous chapter to make RWD more reflective of typical RCT data. Given RCT data will have limited follow-up, I used random sampling to create pseudo-cohorts reflecting a range of sample sizes, follow-up times, and censoring distributions. I assessed model performance using rMSE to understand impacts of these components on extrapolated survival estimates. I also presented impacts on RMST and additionally reported model performance via formal goodness-of-fit (AIC).

6.4.2 Summary of key findings

Evaluation of rMSE revealed largely consistent trends across sample sizes and follow-up time in terms of preferred models. For the uniform right-censoring approach, spline models on the log-odds scale performed best. For both partial and fully distributed censoring approaches, lognormal and loglogistic models performed best. This showed a trend indicating that less complex models performed better where more uncertainty existed in the tails of KM estimates. As expected, preferred models across draws became increasingly stable with increases in sample sizes. Preferred models were similar for follow-up times in terms of classes of models[§§§§§] with the exception of the 36-month follow-up time scenario, where the Gompertz model performed best in two partially distributed censoring scenarios.

Evaluation of differences in RMST among models and scenarios revealed that there was substantial overlap between preferred spline and preferred standard parametric models. Accuracy of RMST

[§§§§§] Given the distributions are very similar for the lognormal and loglogistic[95], I consider them similar classes of models, as well the spline-odds models with varying numbers of internal knots.

estimates decreased with distributed censoring approaches and variability decreased with larger sample sizes. Preferred models identified via AIC were inconsistent with those identified via rMSE.

6.4.3 Interpretation of findings

My finding of spline models performing best when using uniform right-censoring aligns with previous research conducted using a large US cancer registry and a Swedish cancer registry.[117,118] Gray et al. (2021) used SEER data to analyse model performance across four cancers, finding that spline models tended to perform best when right-censoring data at 50%, 35%, and 20% survival using differences in predicted vs observed RMST to select best-performing models.[117] Given RMST estimates in my study follow my main measure of accuracy, rMSE, results can be compared. Specifically for the NSCLC 60-69 years-of-age cohort assessed by Gray et al. (2021), which most closely resembled my cohort given rationale provided in the Chapter 3, spline odds and normal models with 1-3 internal knots performed best via measurement with RMST difference.[117] While I used absolute cutoff times for censoring rather than times as a function of percentage survival as per Gray et al. (2021), results remained largely consistent with my finding that spline odds models with 1 and 3 internal knots performed best. I also argue that absolute follow-up times, as opposed to follow-up times as a function of percentage survival which may or may not reflect typical RCT follow-up, may be more applicable to an HTA setting. My findings also align with those of Chen et al. (2024) who used a Swedish cancer registry using a similar approach to that of Gray et al. (2021).[117,118] Chen et al. (2024) found that, by right-censoring data at 2, 3, and 5 years, and comparing accuracy of survival model extrapolations against 10 year observed survival, spline models performed best. This finding also held for right-censored data at 2, 3, 5, and 10 years used to predict 40 year observed survival.[118]

Scenarios evaluating distributed censoring, however, resulted in less complex models being preferred for extrapolation. This may be explained by the tendency of splines to fit observed data very well, but with the potential risk of overfitting data, which can lead to questionable extrapolations.[29] As distributed censoring increases uncertainty in the tails of KM estimates, which is often seen in RCTs

via relatively wider confidence intervals and larger ‘steps’ in KM estimates, ‘models’ of disease progression may be better represented through standard parametric models than splines. This reasoning is further justified when inspecting spline scenarios, given spline models with fewer knots (i.e. 1 vs 3 internal knots) performed better in scenarios where uncertainty in the tails of KM estimates increased.

While the study showed that censoring distributions had the greatest impact on preferred model selection via rMSE, further evaluation of RMST across models showed that increased follow-up time improved accuracy of estimates. Via boxplots, a clear trend in increased accuracy can be seen as follow-up time moves from 12 to 36 months (i.e. median and IQR values and moving closer to zero). Conversely, while larger sample sizes increased precision as expected, they led to less accurate RMST values. While the finding of increased follow-up resulting in improved accuracy is expected, the result of decreased accuracy with larger sample size seems counterintuitive. This may potentially be explained by parametric survival models being more heavily weighted by early survival trends that include increased numbers at risk. With larger sample sizes, extrapolations may follow early trends more strongly and distinct, later long-term hazards trends, as seen in Chapter 5, may be decreasingly captured in the face of heavily weighted early trends. Additional research, however, would be required to investigate this rationale.

Regarding potential impact on cost-effectiveness estimates, evaluation of RMST interquartile ranges showed substantial overlap between loglogistic, lognormal, and best-performing spline odds models across scenarios. This demonstrates that actual impacts on cost-effectiveness estimates may be limited, in this example, given cost-effectiveness analysis scenarios are typically run for all best-performing models, as is standard practice in HTA. However, potential impacts on decision-making will be case-specific and require evaluation on a case-by-case basis. Further evaluation of RMST showed that certain models, such as the exponential and Gompertz, were generally not suitable across scenarios given the substantial under and overestimation of RMST vs observed estimates and best-performing models. As previously noted, median RMST estimates among models across scenarios, aside from the Gompertz, underestimated RMST, indicating that fitting survival models to relatively

short follow-up times, as typically observed in RCTs, may result in underestimation. This underestimation aligns with previous research validating accuracy of extrapolations fitted to an immuno-oncology trial using data cuts of increasing length.[128] While the Gompertz model performed best for two 36-month follow-up time scenarios, inspection of RMST ranges indicated that estimates were relatively volatile, suggesting that Gompertz models should be carefully evaluated for plausibility if used for extrapolation.

The inconsistency of preferred models selected via AIC vs rMSE illustrates that formal goodness-of-fit over observed data may not be an indicator of appropriateness of models for extrapolation and should not be used in isolation for model selection, in-line with guidance.[28,101]

6.4.4 Study strengths and limitations

Strengths of the study include the use of a large database with a long follow-up period. The use of individual-level data allowed for the various scenarios regarding sample size, follow-up time, and censoring to be conducted, which would not have been possible using aggregate data alone. These scenarios were aimed at making findings applicable to HTA where survival modelling is often performed on RCTs with relatively short follow-up, relatively small sample sizes, and distributed rather than uniform right-censoring. The study thus focussed on accuracy of extrapolations (i.e. projected survival model fits to observed long-term estimates) rather than absolute survival estimates. As adjustment for characteristics such as staging was not required given the aim of the study, the exclusion of these characteristics in the database was not considered to be a significant limitation. Likewise, the potential loss of information with the use of data mapped to the OMOP CDM was not considered to be consequential to findings. There were, however, various limitations in the study that should be addressed.

First, I chose the 60-69 stratum in an attempt to account for heterogeneity, given real-world cohorts are expected to be more heterogenous than RCT cohorts due to strict selection criteria for trial enrolment.[10] This, however, limited age ranges in the cohort and may have impacted long-term survival trends. Particularly, age effects may have changed KM shapes given changes in mortality

rates in the last decades of life. Therefore, the degree and characterisation of heterogeneity of survival estimates between the real-world cohort used in this study and a typical RCT cohort, where age ranges are expected to be broader, is uncertain. Relatedly, my study only included lung cancer which, as presented in Chapter 3, may have relatively lower heterogeneity compared with other cancers. Additionally, in cases where new treatments with novel mechanisms of action, such as immunotherapies, result in markedly different survival distributions vs comparator(s), my study may have limited applicability, requiring investigation on a case-by-case basis. Additional research would be required to determine if findings hold for other cancers or diseases and therapies exhibiting different long-term survival trends.

Second, as visual inspection was not feasible given the number of analyses included in the study, the use of rMSE to measure model performance ensured that averaging, in terms of late overestimation and early overestimation and vice-versa which may occur when using RMST, did not result in selection of erroneously accurate models. Visual inspection of individual log-log and survival plots, however, may have provided further insights, particularly with respect to inspection of observed vs modelled hazards. Relatedly, assessment of samples in which some parametric models did not converge would have been beneficial to understand if observed distributions were potentially invalid and whether eliminating non-converging models introduced bias. As this was not feasible given the large number of draws, this remains a limitation.

Third, while the censoring scenarios were aimed at reflecting ranges of censoring distributions potentially observed in RCTs, limitations remain with the approach taken. Ranges used for distributed-censoring scenarios were chosen subjectively as functions of follow-up time and, within censoring windows, censoring was assumed to take place at random times. Although censoring is assumed to be noninformative in RCTs, in a clinical trial environment censoring patterns can potentially be impacted by factors including efficacy, adverse events, and date of recruitment with respect to data lock times[129,130], therefore not truly being random. Approaches such as plasmode simulation[131], where censoring distributions can be based on patterns observed in select RCTs, may have therefore improved validity of censoring distributions. Further, a simulation study incorporating

a variety of survival distributions and censoring patterns representing numerous diseases and scenarios would increased the generalisability of findings. As my study used a sampling approach using RWD, results are specific to the dataset and cohort analysed, limiting broader applicability. Nonetheless, while my study used simplifying assumptions to produce censoring patterns, it serves as a starting point for understanding impacts on extrapolations.

Fourth, censoring was applied in my study 'looking backwards', in that individuals who had not yet had an event and were not censored at the specified follow-up time were censored at random times retrospectively. While censoring would not occur retrospectively in practice, this approach was used to ensure that the specified number of individuals per scenario remained at the end of the specified follow-up times. Thus, excluding individuals who had died during the censoring windows from retrospective censoring may have limited emulation of censoring distributions expected in a typical RCT. Fifth, sample sizes evaluated in the study were chosen subjectively and, although results were broadly consistent, analysis of additional sample sizes may have provided further insights.

Lastly, the study evaluated only all-cause survival and did not incorporate a relative survival framework, where estimates are decomposed to expected and excess hazards. Previous research has found that, while all-cause vs relative survival showed similar findings for accuracy of 10-year extrapolation, a relative survival framework improved accuracy lifetime extrapolations to 40 years.[118] Incorporating a relative survival framework in the current study may have, therefore, provided additional insights.

Despite the study's various limitations, it served to evaluate transportability of findings regarding survival model selection based on RWD to survival modelling of RCT data. Insights are valuable for understanding the impacts of sample size, follow-up time, and censoring in this context.

6.4.5 Conclusions

In conclusion, in this chapter I demonstrated that, when RWD were adjusted to be more reflective of RCT data in terms of sample size, length of follow-up, and censoring, standard parametric survival models were favoured over spline models. More broadly, I showed that transportability of findings for

preferred survival model selection based on RWD should not be simply assumed. Where suitable RWD exists, the approach used in this chapter may be useful in selecting preferred models fitted to RCT data. Given preferred models identified via AIC were at odds with preferred models identified via rMSE alone, my study suggests that goodness-of-fit criteria should not be used in isolation to identify preferred survival models for extrapolation.

7 Discussion

In this final chapter of the thesis, I conclude with an overarching discussion of key findings, implications for decision-making, areas for future research, and recommendations.

7.1 Summary of key findings

My thesis investigated approaches for using real-world TTE data to supplement RCT data in HTA. I focussed on cases where only aggregate data may be available or adjustment of individual-level RWD to represent a target RCT population may not be feasible. The thesis was undertaken with the aim of being directly applicable to cost-effectiveness analysis used in HTA. In Chapter 2, I found that previous studies aiming to generalise cost-effectiveness estimates most often used RWD in place of RCT data to specify baseline risks for this purpose. This aligns with stakeholder recommendations for using RWD to capture natural history as a potential *"Easy-Win"* for integrating RWD in HTA, with a view that generalisability of trial data is a high-priority area.[38] The studies identified in the literature review that incorporated real-world baseline TTE data in cost-effectiveness modelling, however, lacked justification for approaches used. Although these studies were predominantly designed to confirm cost-effectiveness estimates rather than explore methodological approaches, the lack of justification may reflect the lack of practical guidance on integrating RWD alongside RCT evidence in HTA. As evident in later chapters of the thesis, integrating real-world TTE baseline data in cost-effectiveness models can potentially become challenging for numerous reasons, and therefore needs to be performed transparently and include thorough exploration of uncertainties inherent with the use of RWD, as per previous recommendations.[27] Thus, based on my literature review and absence of existing guidance, I concluded that further research into the impacts of various methodological approaches for integrating real-world TTE baseline data in HTA was warranted.

After describing data sources and common survival methods used throughout the thesis in Chapter 3, in Chapter 4 I illustrated challenges associated with integrating aggregate real-world baseline TTE data when faced with complex hazards. I showed that cost-effectiveness estimates can be substantially

impacted by approaches employed. Substituting RCT with real-world baseline risks from a well-designed observational study and applying relative effects from an RCT may, at first glance, seem straightforward. However, in keeping with best practices for survival modelling for HTA, hazards between RCT and real-world baselines and treatment effects need to be considered. Typically, when using RCT data alone, where the proportional hazards assumption holds, a hazard ratio can be applied and, where it does not, treatment arms can be modelled independently to capture non-proportional hazards and preserve treatment effects as observed. However, where treatment effects are applied to an external independent distribution, independent modelling of treatment arms is not possible, thus more flexible, and potentially complex approaches must be used. The chapter highlighted that, while thoroughly exploring impacts of these approaches on cost-effectiveness estimates may entail a substantial amount of work, not doing this can potentially lead to biased estimates. This was illustrated by the substantially different cost-effectiveness estimates produced among approaches for capturing real-world reference treatment/baseline risks. Furthermore, although different approaches for treatment effect specifications resulted in relatively smaller differences in cost-effectiveness estimates, I argued that exploring potential impacts is required on a case-by-case basis if trust in estimates is to be gained for use in decision-making.

In Chapter 5, I outlined an approach for using RWD to inform RCT-based modelling approaches, where direct application of RWD in a cost-effectiveness model may be considered inappropriate. Given the level of detail available in the dataset used, RWD could not be matched to the target RCT population in terms of characteristics such as staging or specific lung cancer subtype, and thus emulation of the target trial population was not possible. One could argue that, given the lack of detail available in the dataset, the RWD in this case should be dismissed given the degree of uncertainty regarding representativeness of the target trial population. However, by further assessing the data for broader representativeness while acknowledging limitations, I demonstrated that such RWD may be useful in the absence of other evidence to inform long-term extrapolation. Where the only other evidence available to support long-term survival extrapolation is general population background mortality or expert opinion, I argue that limited RWD is a step forward in informing plausible

estimates. Using a motivating example, the chapter showed how the use of RWD may have been helpful for decision-making in a previous HTA where extrapolated survival estimates were highly contested among stakeholders.

In Chapter 6, I explored the impact of censoring patterns, length of follow-up, and sample size when using RWD to inform survival model selection for extrapolating RCT data. This was motivated by Chapter 5 as well as calls for research in this area.[117] By making RWD more reflective of what would be expected in a typical RCT, it was found that relatively less complex standard survival models may be more appropriate for extrapolation, compared with more complex spline models. This finding was predominantly driven by censoring patterns rather than sample size or follow-up time, indicating that increased uncertainty typically observed in the tails of RCT KM estimates should be considered when using RWD to inform model selection. Although preferred model selection differed by introduced censoring distribution scenarios, assessment of impact on RMST, a key metric used to estimate cost-effectiveness, showed substantial overlap among best-performing models. This suggests that actual impacts on cost-effectiveness may be minimal. Nonetheless, taking the cost-effectiveness results from Chapter 5 as an example, a 46% difference in ICERs between the top two performing models may well be important where new drugs are priced close to willingness-to-pay thresholds. Acknowledging that evaluation would need to be performed on a case-by-case basis, I concluded that differences between real-world and RCT data, such as sample size, length of follow-up, and censoring distributions, should be considered when using RWD to inform survival model selection.

7.2 Implications for decision-making

In the context of developing economic models to inform decision-making, Briggs et al. (2006) state that researchers are required *"to incorporate all appropriate evidence into the analysis"*.[4] Drummond et al. (2015) mirror this, stating that *"A fundamental tenet of evidence-based decision making is the need to use all relevant evidence in an analysis undertaken to inform decisions and policy."*.[3] Given the nuances involved in generating evidence from RWD, what qualifies as appropriate or relevant

evidence is necessarily subjective and should be evaluated on a case-by-case basis. As my thesis explored practical approaches for the use of relatively limited RWD to inform baseline survival using various strategies, there are numerous implications for those involved in decision-making processes.

For analysts, assuming a well-conducted observational study specific to the population under investigation in an HTA has been performed, integrating this evidence in a cost-effectiveness model may still be complex if only aggregate data are available. Potential complexities, however, can be overcome via various survival modelling approaches, as illustrated in Chapter 4, and should not be viewed as a barrier for including RWD to supplement RCT-based estimates. By using systematic approaches, exploring uncertainties, and assessing impacts on cost-effectiveness estimates, trust in RWD to inform decision-making can, arguably, be gained. Further, in situations such as managed-access agreements or 'living-HTA', where cost-effectiveness estimates are iteratively updated and rely on RWD[110], approaches such as those presented in the thesis for handling complex data may be required. My thesis thus motivates the use of RWD, in spite of complexities, for formally assessing external validity as well as iterative updates of cost-effectiveness estimates to inform decision-making.

Further, where RWD are not appropriate for direct application in cost-effectiveness modelling, RWD can still be used indicatively by analysts to assess suitability of extrapolations for target populations. While a relatively crude example of this is the use of absolute survival percentages, the thesis illustrated that additional analyses, such as assessing hazards trends in longer-term RWD, can be used to inform survival trends expected beyond follow-up periods typically observed in RCTs. By fully presenting limitations in particular real-world datasets and making the best use of available information, valuable information can be utilised to aid in model selection which may impact cost-effectiveness estimates and therefore decision-making in HTA.

For decision makers, in the face of limited RWD, it could be argued that dismissing such evidence for consideration in HTA is a safe option. However, given RCTs typically include international populations with characteristics that may not fully represent individuals receiving treatment in routine clinical practice, one could consider that basing decisions on RCT data alone may be questionable.

Given all RWD, to a certain extent, is limited, by accepting that evidence generated from RWD is not a replacement for RCT evidence, RWD can potentially be leveraged to a greater extent and used appropriately to inform decision-making. Given HTA submissions should adequately acknowledge and explore the potential impacts of inherent limitations in incorporated data, I argue that the risk of using RWD to inform HTA is relatively small. Excluding RWD on grounds of poor-quality evidence may, in fact, hold greater risks for decision-making on the basis of omitted evidence.

For companies, payers, and perhaps research organisations, increasing accessibility to real-world datasets and evidence would be a great step forward in increasing transparency and providing common 'points of truth' for stakeholders. My thesis investigated both aggregate and individual-level RWD, neither of which allowed for extensive adjustment. This led to a substantial amount of work being required in attempts to address uncertainties and limitations in the data. The development of validated, publicly available platforms allowing for adequate adjustment without the need to access individual-level data, would surely be beneficial for stakeholders. One could see a future where predetermined baseline risks specific to the target population of interest can routinely be used in HTA, decreasing uncertainty for all parties involved in the decision-making process. This would have the added benefit of making cost-effectiveness estimates more directly applicable to the local target population for which decisions are being made. In turn, this could lead to more efficient HTA processes and, ultimately, better-informed decision-making.

Overarchingly, while it could be argued that the approaches presented in the thesis could entail a substantial amount of effort from analysts, the sensitivity of cost-effectiveness estimates to approaches highlights the importance of such efforts for decision-making. Currently, the added benefit of utilising RWD to supplement RCT data may be questioned due to limited understanding of methods, lack of guidance, and uncertainty regarding potential impacts on estimates. The thesis highlighted the potential consequences of inappropriately integrating or excluding RWD to supplement RCT data in economic evaluation, which can ultimately result in significant cost to the healthcare system and/or restricted access to new interventions, ultimately affecting patient care.

In summary, until the integration of RWD into economic evaluation is widely accepted and standardised, the methodological contributions of this thesis serve as a resource for understanding the potential impacts of presented approaches for using real-world baseline TTE data to supplement RCT data for HTA. The thesis illustrates that, by using systematic, transparent, and methodologically appropriate approaches for including relatively limited RWD in HTA, trust in using such data by stakeholders can potentially be gained.

7.3 Areas for future research

The thesis presented a range of approaches and considerations that should be acknowledged when using real-world TTE baseline data to supplement RCT data for HTA. In situations where RWD may lack certain characteristics required to adjust individual-level data or exist only in aggregate form, these approaches serve as potential paths for appropriately integrating such data. Throughout my research, I identified several areas for future research, broadly covering impacts of methodological approaches for integrating real-world baseline TTE data in cost-effectiveness analysis.

First, in my thesis I described the need to apply relative effects from clinical trials to real-world baselines for new technologies at the time of HTA, given the absence of data on real-world effects. This, however, can lead to biased estimates when using probabilistic analysis.[64] This is because the covariance between relative effects and baseline estimates is not preserved, as would be the case with joint modelling, where baseline risk and treatment effects simultaneously. Further, relative effect estimates from one parametric model should, technically, not be applied to a separate independent model[28,64], which would be required when combining RCT and RWD as per the approach presented in Chapter 4. While this appears to be an ongoing 'accepted' limitation in HTA[64], and has been acknowledged in previous guidance[28], further research into methods for better characterising and/or understanding impacts of this limitation on cost-effectiveness estimates is warranted.

Second, in my thesis I proposed approaches for assessing RWD-based estimates as a supplement to, rather than a replacement for, RCT-based estimates. If such approaches would be deemed acceptable

for use in HTA, the next logical step would be determining the extent to which this supplemental information should be used for decision making. Synthesising real-world with RCT-based cost-effectiveness estimates is therefore an important consideration. Methods for deriving a single ICER using independent sources include cost-effectiveness model averaging, where various scenarios are averaged, Bayesian approaches, where multiple data sources are synthesised[132], and using RCT data which has been adjusted to reflect target population characteristics[50], among others. Differences in cost-effectiveness estimates using various methods is warranted to understand sensitivity to approaches used. This is an important area for future research if scenarios using RWD to supplement RCT estimates are formally integrated in HTA.

Third, the thesis included the use of RWD to inform long-term survival trends for HTA. In Chapter 5, I assessed suitability of RWD to inform survival trends for a particular population and concluded that such analysis would be required on a case-by-case basis if used for HTA. To expand on this, future research could investigate heterogeneity of long-term hazards trends among various subpopulations and disease areas. A body of literature assessing differences in long-term hazards trends across populations would be useful in understanding requirements for RWD for informing extrapolation. On the one hand, if trends were relatively homogenous after a certain time point, one could, for example, argue that real-world survival for particular diseases could be used to inform long-term estimates across subpopulations. Estimates could consequently be applied directly in cost-effectiveness models in the same way that background mortality is often applied, serving as a bridge between short-term RCT data and long-term general population background mortality. Conversely, if heterogeneity was substantial across subpopulations, it would indicate that population-specific data are required on a case-by-case basis. Relatedly, research investigating generalisability of hazards trends across counties would be useful for instances where country-specific data, for example, is unavailable for decision-making. This work could further be expanded to include investigation of statistical cure modelling. Given the low numbers at risk at the tails of typical RCT KM estimates, assuming statistical cure based on RCT evidence may be questionable.[29] If statistical cure were estimated using real-world

datasets with large sample sizes and adequate follow-up not typical of RCTs for making such claims, the often contested cure-point models used in HTAs[133] might have a stronger footing.

Furthermore, if access to sufficient RWD were available, investigating differences in short-term hazards would be useful. Given the important of proportional hazards involved with survival analysis, complex hazards seen in RCT KM estimates could, in part, be due to relatively low sample sizes. Comparing RCT hazards vs sufficiently matched real-world short-term hazards would thus be informative regarding potential data artefact. This could lead to relaxation of the proportional hazards assumption (i.e. assuming proportional hazards), for example, where complex hazards seen via RCT KM estimates are not expected to be representative of true hazards in average populations.

Fourth, as presented in Chapter 6, increased uncertainty in the tails of KM estimates led to standard, less complex models being preferred over splines for long-term extrapolations. To better understand the impact of uncertainty, future research could investigate survival model performance as a function of uncertainty in the tails of KM estimates. While my study artificially introduced uncertainty through distributed censoring, sample size, and follow-up time, investigating accuracy as a function of confidence interval thresholds, for example, may provide more specific insights.

Fifth, as observed survival in Chapter 6 was best captured via standard distributions, more complex hazards, as were seen in Chapter 4, may require spline or other flexible models to be fitted over, or to the early section of the observed period (i.e. where standard parametric models cannot capture observed hazards). Based on findings from Chapter 6 that splines may potentially lead to questionable extrapolations, this situation may become problematic where exclusion of early data (i.e. if fitting piecewise models) results in loss of information for modelling long-term trends.[31] Future research could thus investigate trade-offs between prioritising observed vs extrapolated estimates. This research could investigate whether hybrid models, such as the direct use of KM estimates or flexible models over observed periods appended with a standard parametric model, fitted to all or a segment of the observed data, may be suitable.

Lastly, as acknowledged in previous guidance[29], Bayesian approaches hold promise for integrating RWD in HTA. Although my thesis did not focus on Bayesian approaches, future research on methods for application in HTA seems warranted.

7.4 Recommendations

There are overarching recommendations stemming from the thesis. First, when aiming to generalise cost-effectiveness estimates using RWD, it is important to provide comprehensive reporting to avoid making potentially erroneous claims by neglecting uncertainties and limitations. Clear guidance or checklists tailored explicitly for this purpose would be beneficial.

Second, echoing previous calls[11], generalisability of cost-effectiveness estimates based on RCT data should explicitly be analysed in HTA, given potential impacts on decision-making. Such analyses should be conducted using appropriate methods if they are to form the evidence base for decision-making. Current guidance, such as the Consolidated Health Economic Evaluation Reporting Standards 2022 Statement[134], should consider including generalisability in the methods, in addition to the discussion, section of the checklist.

Third, in cases where RWD is considered insufficient for direct use in cost-effectiveness modelling due to a lack of characteristics required for extensive adjustment approaches, such RWD should still be acknowledged as a potential data source of information to form the evidence base for decision-making. Careful consideration of limitations should be thoroughly considered to ensure false certainty is not introduced into the decision-making process.

Lastly, differences between real-world and RCT data (e.g. sample size, length of follow, censoring) should be fully acknowledged, and transportability of methods between these different sources of data should not be assumed. Approaches deemed suitable for analysing RWD may not necessarily be suitable for RCT analysis, and vice versa, due to variations in information and data included between datasets.

7.5 Conclusions

The thesis had the overarching aim of investigating approaches for using RWD to supplement RCT data for use in HTA, specifically when considering the use of baseline TTE data for cost-effectiveness modelling. A review of the literature identified various parameters and methodologies for generalising cost-effectiveness estimates using RWD, with baseline risk being the most commonly generalised parameter. The review's findings suggested the need for methodological guidance to integrating real-world baseline TTE data due to the lack systematic or consistent approaches identified in the literature. When investigating approaches for integrating real-world baselines with trial-derived relative treatment effects, I found that methods for capturing baseline survival substantially impacted cost-effectiveness estimates, necessitating a balance between model complexity and suitability for decision-making. When investigating application of real-world overall survival for HTA, where estimates could not be fully adjusted to match a target trial population, I found that such RWD can inform extrapolation and, in turn, decision-making by providing valuable insights regarding appropriateness of survival models used for extrapolation. Suitability of such data for use in HTA should, however, be evaluated on a case-by-case basis. When investigating the impact of introduced censoring patterns, sample sizes, and follow-up times to make RWD more reflective of data typically observed in trials, I found that standard parametric models generally performed better than spline-based models in terms of predictive accuracy of extrapolations. This highlighted that transportability of methods between real-world and RCT data should not be assumed. The thesis has thus contributed to the evidence base by providing guidance on approaches for integrating RWD with clinical trial data and illustrating potential impacts on cost-effectiveness estimates. The thesis motivates the use of RWD in HTA as a valuable source of evidence to inform decision-making.

8 Bibliography

1 OECD. Fiscal Sustainability of Health Systems. Bridging Health and Finance Perspectives. 2015.

2 4 Economic evaluation | NICE health technology evaluations: the manual | Guidance | NICE. https://www.nice.org.uk/process/pmg36/chapter/economic-evaluation (accessed March 8, 2022).

3 Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes, Fourth edition. Oxford: Oxford University Press, 2015.

4 Briggs A, Claxton K, Schulpher M. Decision modelling for health economic evaluation. Oxford: Oxford University Press, 2006.

5 National Institute for Health and Care Excellence. Real world evidence framework. 2022.

6 Canada's Drug and Health Technology Agency. Guidance for reporting Real-World Evidence. 2023.

7 Hampson G, Towse A, Dreitlein WB, Henshall C, Pearson SD. Real-world evidence for coverage decisions: Opportunities and challenges. *J Comp Eff Res* 2018; **7**: 1133–43.

8 Makady A, van Veelen A, Jonsson P, *et al.* Using Real-World Data in Health Technology Assessment (HTA) Practice: A Comparative Study of Five HTA Agencies. *Pharmacoeconomics* 2018; **36**: 359–68.

9 Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *NPJ Digit Med* 2020; **3**: 67.

10 Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015. DOI:10.1186/s13063-015-1023-4.

11 Van Staa TP, Leufkens HG, Zhang B, Smeeth L. A comparison of cost effectiveness using data from randomized trials or actual clinical practice: Selective Cox-2 inhibitors as an example. *PLoS Med* 2009; **6**. DOI:10.1371/journal.pmed.1000194.

12 Gill JL, Kanavos PG. Real-world evidence in Europe – An analysis. *Journal of Health Policy* 2018; : 4–9.

13 Bell H, Wailoo AJ, Hernandez M, *et al.* The use of real world data for the estimation of treatment effects in NICE decision making. 2016. http://www.nicedsu.org.uk.

14 Makady A, Ham R ten, de Boer A, Hillege H, Klungel O, Goettsch W. Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value in Health* 2017; **20**: 520–32.

15 Bullement A, Podkonjak T, Robinson MJ, *et al.* Real-world evidence use in assessments of cancer drugs by NICE. *Int J Technol Assess Health Care* 2020; **36**: 388–94.

16 Makady A, Stegenga H, Ciaglia A, *et al.* Practical implications of using real-world evidence (RWE) in comparative effectiveness research: learnings from IMI-GetReal. *J Comp Eff Res* 2017; **6**: 485–90.

17 Kearns B, Poku E, Wailoo A, Avey B, Buckley C, Taylor M. Centre for Health Technology Evaluation Methods Guide Update: Consultation Response Report. Report by the Decision Support Unit and External Assessment Centre. 2021.

18 Faria R, Alava MH, Manca A, Wailoo AJ. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness for Technology Appraisal: Methods for comparative individual patient data. 2015 www.nicedsu.org.uk.

19 Davies A, Briggs A, Schneider J, *et al.* The Ends Justify the Mean: Outcome Measures for Estimating the Value of New Cancer Therapies. *Health Outcomes Res Med* 2012; **3**: e25–36.

20 Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence Synthesis for Decision Making 5: The Baseline Natural History Model. *Medical Decision Making* 2013; **33**: 657–70.

21 Woods B, Sideris E, Palmer S, Latimer N, Soares M. NICE DSU Technical Support Document 19. Partitioned Survival Analysis for Decision Modelling in Health Care: A Critical Review. 2017 www.nicedsu.org.uk.

22 Canadian Agency for Drugs and Technologies in Health. CADTH Methods and Guidelines. Guidelines for the Economic Evaluation of Health Technologies: Canada. 4th Edition. 2017.

23 Zorginstituut Nederland. Guideline for economic evaluations in healthcare. 2016.

24 Pharmaceutical Benefits Advisory Committee. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee. Version 5.0. 2016.

25 National Institute for Health and Care Excellence. NICE health technology evaluations: the manual. Process and methods. 2022 www.nice.org.uk/process/pmg36.

26 Latimer NR, Adler AI. Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ medicine* 2022; **1**: e000094.

27 Jackson C, Stevens J, Ren S, *et al.* Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods. *Medical Decision Making* 2017; **37**: 377–90.

28 Latimer N. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials -extrapolation with patient-level data. 2011 http://www.nicedsu.org.uk.

29 Rutherford MJ, Lambert PC, Sweeting MJ, *et al.* NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis. 2020 www.nicedsu.org.uk.

30 Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: An alternative approach. *Medical Decision Making* 2014. DOI:10.1177/0272989X13497998.

31 Latimer NR. Response to 'survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach' by Bagust and Beale. *Med Decis Making* 2014; **34**: 279–82.

32 National Institute for Health and Care Excellence. Atezolizumab with carboplatin and etoposide for untreated extensive-stage small-cell lung cancer [ID1504]. Technology Appraisal Guidance[TA638]. Committee Papers. 2020.

33 Bullement A, Latimer NR, Bell Gorrod H. Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study. *Value Health* 2019; **22**: 276–83.

34 National Institute for Health and Care Excellence. Nivolumab for previously treated locally advanced or metastatic squamous non-small-cell lung cancer. Committee Papers. (Draft guidance: 1 - TA483). 2015.

35 Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information. *Med Decis Making* 2017; **37**: 353–66.

36 European Health Data & Evidence Network (EHDEN). European health data evidence network [WWW document]. 2018. https://www.ehden.eu/ (accessed Feb 19, 2024).

37 Claire R, Elvidge J, Read C, *et al.* RWD147 Demonstrating the Application of Real-World Evidence for Health Technology Assessment Using a Federated Data Network. *Value in Health* 2022; **25**: S478.

38 Claire R, Elvidge J, Hanif S, *et al.* Advancing the use of real world evidence in health technology assessment: insights from a multi-stakeholder workshop. *Front Pharmacol* 2023; **14**: 1289365.

39 Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 2018; **18**. DOI:10.1186/s12874-018-0611-x.

40 Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016; **5**. DOI:10.1186/S13643-016-0384-4.

41 de Pouvourville G, Blin P, Karam P. The contribution of real-world evidence to cost-effectiveness analysis: case study of Dabigatran etexilate in France. *European Journal of Health Economics* 2020; **21**: 235–49.

42 Hollier-Hann G, Curry A, Onishchenko K, *et al.* Updated cost-effectiveness analysis of onabotulinumtoxinA for the prevention of headache in adults with chronic migraine who have previously received three or more preventive treatments in the UK. *J Med Econ* 2020; **23**: 113–23.

43 de Jong LA, Groeneveld J, Stevanovic J, *et al.* Cost-effectiveness of apixaban compared to other anticoagulants in patients with atrial fibrillation in the real-world and trial settings. *PLoS One* 2019; **14**: 1–17.

44 Desai M, Bentley A, Keck WA, Haag T, Taylor RS, Dakin H. Cooled radiofrequency ablation of the genicular nerves for chronic pain due to osteoarthritis of the knee: A cost-effectiveness analysis based on trial data. *BMC Musculoskelet Disord* 2019; **20**. DOI:10.1186/s12891-019-2681-2.

45 Wan W, Skandari MR, Minc A, *et al.* Cost-effectiveness of continuous glucose monitoring for adults with type 1 diabetes compared with self-monitoring of blood glucose: The DIAMOND randomized trial. *Diabetes Care* 2018; **41**: 1227–34.

46 Seferina SC, Ramaekers BLT, de Boer M, *et al.* Cost and cost-effectiveness of adjuvant trastuzumab in the real world setting: A study of the Southeast Netherlands Breast Cancer Consortium. *Oncotarget* 2017; **8**: 79223–33.

47 Yeung K, Li M, Carlson JJ. Using Performance-Based Risk-Sharing Arrangements to Address Uncertainty in Indication-Based Pricing. *J Manag Care Spec Pharm* 2017; **23**: 1010–5.

48 van Kampen RJW, Ramaekers BLT, Lobbezoo DJA, *et al.* Real-world and trial-based cost-effectiveness analysis of bevacizumab in HER2-negative metastatic breast cancer patients: a study of the Southeast Netherlands Breast Cancer Consortium. *Eur J Cancer* 2017; **79**: 238–46.

49 van der Linden N, van Gils CWM, Pescott CP, Buter J, Vergeer MR, Groot CAU de. Real-world cost-effectiveness of cetuximab in locally advanced squamous cell carcinoma of the head and neck. *European Archives of Oto-Rhino-Laryngology* 2015; **272**: 2007–16.

50 Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society* 2015; **178**: 757–78.

51 Blommestein HM, Issa DE, Pompen M, *et al.* Cost-effectiveness of rituximab as maintenance treatment for relapsed follicular lymphoma: Results of a population-based study. *Eur J Haematol* 2014; **92**: 398–406.

52 Morais J, Aguiar C, McLeod E, Chatzitheofilou I, Fonseca Santos I, Pereira S. Cost-effectiveness of rivaroxaban for stroke prevention in atrial fibrillation in the Portuguese setting. *Revista Portuguesa de Cardiologia* 2014; **33**: 535–44.

53 Chevalier J, Delaitre O, Hammès F, De Pouvourville G. Cost-effectiveness of dabigatran versus vitamin K antagonists for the prevention of stroke in patients with atrial fibrillation: A French payer perspective. *Arch Cardiovasc Dis* 2014; **107**: 381–90.

54 Foglia E, Bonfanti P, Rizzardini G, *et al.* Cost-utility analysis of lopinavir/ritonavir versus atazanavir + ritonavir administered as first-line therapy for the treatment of HIV infection in Italy: from randomised trial to real world. *PLoS One* 2013; **8**. DOI:10.1371/JOURNAL.PONE.0057777.

55 Van Gils CWM, De Groot S, Redekop WK, Koopman M, Punt CJA, Uyl-De Groot CA. Real-world cost-effectiveness of oxaliplatin in stage III colon cancer: a synthesis of clinical trial and daily practice evidence. *Pharmacoeconomics* 2013; **31**: 703–18.

56 Wouters H, Thijs V, Annemans L. Cost-effectiveness of dabigatran etexilate in the prevention of stroke and systemic embolism in patients with atrial fibrillation in Belgium. *J Med Econ* 2013; **16**: 407–14.

57 Brandes DW, Raimundo K, Agashivala N, Kim E. Implications of real-world adherence on cost-effectiveness analysis in multiple sclerosis. *J Med Econ* 2013; **16**: 547–51.

58 Bellows BK, Dahal A, Jiao T, Biskupiak J. A cost-utility analysis of pregabalin versus duloxetine for the treatment of painful diabetic neuropathy. *J Pain Palliat Care Pharmacother* 2012; **26**: 153–64.

59 Gillett M, Dallosso HM, Dixon S, *et al.* Delivering the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: Cost effectiveness analysis. *BMJ (Online)* 2010; **341**. DOI:10.1136/bmj.c4093.

60 Sorensen S V., Dewilde S, Singer DE, Goldhaber SZ, Monz BU, Plumb JM. Cost-effectiveness of warfarin: Trial versus real-world stroke prevention in atrial fibrillation. *Am Heart J* 2009; **157**: 1064–73.

61 Annemans L, Strens D, Lox E, Petit C, Malonne H. Cost-effectiveness analysis of aprepitant in the prevention of chemotherapy-induced nausea and vomiting in Belgium. *Support Care Cancer* 2008; **16**: 905–15.

62 Farahani P, Levine M, Goeree R. A comparison between integrating clinical practice setting and randomized controlled trial setting into economic evaluation models of therapeutics. *J Eval Clin Pract* 2006; **12**: 463–70.

63 Reed SD, Dillingham PW, Briggs AH, Veenstra DL, Sullivan SD. A Bayesian approach to aid in formulary decision making: incorporating institution-specific cost-effectiveness data with clinical trial results. *Med Decis Making* 2003; **23**: 252–64.

64 Guyot P, Welton NJ, Ouwens MJNM, Ades AE. Survival time outcomes in randomized, controlled trials and meta-analyses: The parallel universes of efficacy and cost-effectiveness. *Value in Health* 2011; **14**: 640–6.

65 University of York. Centre for Reviews and Dissemination. https://www.crd.york.ac.uk/CRDWeb/SearchPage.asp (accessed Jan 14, 2025).

66 The International Network of Agencies for Health Technology Assessment. International HTA database. https://database.inahta.org/ (accessed Jan 14, 2025).

67 Tufts Medical Center. Centre for the Evaluation of Value and Risk in Health. CEA Registry - Center for the Evaluation of Value and Risk in Health. https://cevr.tuftsmedicalcenter.org/databases/cea-registry (accessed Jan 14, 2025).

68 Van Der Linden N, Van Gils CWM, Pescott CP, Buter J, Uyl-De Groot CA. Cetuximab in locally advanced squamous cell carcinoma of the head and neck: Generalizability of EMR 062202-006 trial results. *European Archives of Oto-Rhino-Laryngology* 2014; **271**: 1673–8.

69 Bonner JA, Harari PM, Giralt J, *et al.* Radiotherapy plus Cetuximab for Squamous-Cell Carcinoma of the Head and Neck. *New England Journal of Medicine* 2006; **354**: 567–78.

70 Rohatgi A. WebPlotDigitizer. 2022; published online Sept. https://automeris.io/WebPlotDigitizer.

71 Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol* 2011; **11**: 1–14.

72 Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012. DOI:10.1186/1471-2288-12-9.

73 Brahmer J, Reckamp KL, Baas P, *et al.* Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer. *New England Journal of Medicine* 2015; **373**: 123–35.

74 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.

75 Clinical Practice Research Datalink. Release Notes: CPRD GOLD July 2020. 2020.

76 Clinical Practice Research Datalink. Release Notes: CPRD GOLD July 2022. 2022.

77 Voss EA, Makadia R, Matcho A, *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association* 2015; **22**: 553–64.

78 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015; **216**: 574–8.

79 Schuemie M, Suchard M. OhdsiRTools: Tools Used by Observational Health Data Science and Informatics (OHDSI). 2022 http://ohdsi.github.io/OhdsiRTools/ (accessed Feb 15, 2024).

80 Gilbert J, Rao G, Schuemie M, Ryan P, Weaver J. CohortDiagnostics: Diagnostics for OHDSI Cohorts. 2023. https://ohdsi.github.io/CohortDiagnostics.

81 Sena AG, Schuemie MJ. CohortGenerator: An R package for Cohort Generation against the OMOP CDM. 2023. https://ohdsi.github.io/CohortGenerator (accessed May 17, 2024).

82 Black A, Gorbachev A, Burn E, Sabate MS. CDMConnector: Connect to an OMOP Common Data Model. 2024. https://github.com/darwin-eu/CDMConnector (accessed May 17, 2024).

83 Newby D. Incidence, Prevalence and Survival of Lung Cancer in the United Kingdom between 2000 and 2021. Web app companion to Corby G, Barclay NL, Tan EH, et al. Incidence, prevalence, and survival of lung cancer in the United Kingdom from 2000-2001: a population-based cohort study. medRxiv (2024). 2024. https://dpa-pde-oxford.shinyapps.io/EHDENCancerIncPrevCohortDiagShiny/ (accessed Aug 6, 2024).

84 Strongman H, Williams R, Bhaskaran K. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? a concordance and validation study using linked English electronic health records data. *BMJ Open* 2020; **10**. DOI:10.1136/BMJOPEN-2020-037719.

85 van Veelen A, Abtahi S, Souverein P, *et al.* Characteristics of patients with lung cancer in clinical practice and their potential eligibility for clinical trials evaluating tyrosine kinase inhibitors or immune checkpoint inhibitors. *Cancer Epidemiol* 2022; **78**. DOI:10.1016/j.canep.2022.102149.

86 Corby G, Barclay NL, Tan EH, *et al.* Incidence, prevalence, and survival of lung cancer in the United Kingdom from 2000-2021: a population-based cohort study. *medRxiv* 2024; **4**: 2024.02.26.24303406.

87 Cancer Research UK. Lung cancer. Stages and types of lung cancer. Types of lung cancer. https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types (accessed Sept 17, 2023).

88 Cancer Research UK. Lunch cancer statistics. Lung cancer incidence. Early diagnosis data hub. Incidence by stage. Proportion diagnosed by state (England, 2020). https://crukcancerintelligence.shinyapps.io/EarlyDiagnosis/ (accessed Sept 17, 2023).

89 R Core Team. A language and environment for statistical computing. . Vienna, 2019 https://www.R-project.org/ (accessed Feb 15, 2024).

90 Therneau T. A Package for Survival Analysis in R. R package version 3.5-5, https://CRAN.R-project.org/package=survival. 2023.

91 Jackson CH. flexsurv: A Platform for Parametric Survival Modeling in R. *J Stat Softw* 2016; **70**. DOI:10.18637/jss.v070.i08.

92 Wickham H, Averick M, Bryan J, *et al.* Welcome to the Tidyverse. *J Open Source Softw* 2019; **4**: 1686.

93 Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. 2021 https://CRAN.R-project.org/package=survminer (accessed Feb 15, 2024).

94 Hess Kenneth, Gentleman R. Hazard Function Estimation in Survival Analysis. 2021 https://CRAN.R-project.org/package=muhaz (accessed Feb 15, 2024).

95 Collett D. Modelling survival data in medical research. P.157-159. CRC press, 2023.

96 Ouwens MJNM, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Res Synth Methods* 2010. DOI:10.1002/jrsm.25.

97 Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. *Stata Journal* 2011.

98 Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Comput Simul* 2015; **85**: 777–93.

99 Harrell FE. Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. 2001 DOI:10.1007/978-1-4757-3462-1.

100 Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network Meta-Analysis for Decision Making. 2018 DOI:10.1002/9781118951651.

101 Palmer S, Borget I, Friede T, *et al.* A Guide to Selecting Flexible Survival Models to Inform Economic Evaluations of Cancer Immunotherapies. *Value Health* 2023; **26**: 185–92.

102 Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. 1998. DOI:10.1007/978-1-4612-1694-0_15.

103 Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics* 1978. DOI:10.1214/aos/1176344136.

104 Michiels S, Le Maître A, Buyse M, *et al.* Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data. *Lancet Oncol* 2009; **10**: 341–50.

105 Black CM, Keeping S, Mojebi A, *et al.* Correlation Between Early Time-to-Event Outcomes and Overall Survival in Patients With Locally Advanced Head and Neck Squamous Cell Carcinoma Receiving Definitive Chemoradiation Therapy: Systematic Review and Meta-Analysis. *Front Oncol* 2022; **12**: 1.

106 Garattini L, Padula A. Dutch guidelines for economic evaluation: 'from good to better' in theory but further away from pharmaceuticals in practice? *J R Soc Med* 2017; **110**: 98.

107 Vardell E. Global Health Observatory Data Repository. *Med Ref Serv Q* 2020. DOI:10.1080/02763869.2019.1693231.

108 NHS Northern and Yorkshire Regional Drug and Therapeutics Centre (RDTC) and the Centre for Health Economics (CHE) U of York. Cetuximab for the treatment of locally advanced squamous cell cancer of the head and neck. Technology appraisal guidance [TA145]. Head and neck cancer - cetuximab: evidence review group report. 2008.

109 IQVIA. Impact of RWE on HTA Decision-making. 2022.

110 Thokala P, Srivastava T, Smith R, *et al.* Living Health Technology Assessment: Issues, Challenges and Opportunities. *Pharmacoeconomics* 2023; **41**: 227–37.

111 Hatswell AJ, Bullement A, Briggs A, Paulden M, Stevenson MD. Probabilistic Sensitivity Analysis in Cost-Effectiveness Models: Determining Model Convergence in Cohort Models. *Pharmacoeconomics* 2018; **36**: 1421–6.

112 Wang X, Adamson BJ, Briggs A, *et al.* Approaches for Enhanced Extrapolation of Long-Term Survival Outcomes Using Electronic Health Records of Patients With Cancer. *Value in Health* 2022; **25**: 230–7.

113 Wooldridge J. Introductory econometrics. A modern approach. 5th ed. Australia: South-Western, 2013 DOI:10.1016/j.jconhyd.2010.08.009.

114 Royston P, Parmar MKB. 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 Med Res Methodol* 2013; **13**: 1–15.

115 Ouwens MJNM, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating Lifetime Benefits Associated with Immuno-Oncology Therapies: Challenges and Approaches for Overall Survival Extrapolations. *Pharmacoeconomics* 2019; **37**: 1129–38.

116 Bullement A, Latimer NR, Bell Gorrod H. Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study. *Value in Health* 2019; **22**: 276–83.

117 Gray J, Sullivan T, Latimer NR, *et al.* Extrapolation of Survival Curves Using Standard Parametric Models and Flexible Parametric Spline Models: Comparisons in Large Registry Cohorts with Advanced Cancer. *Med Decis Making* 2021; **41**: 179–93.

118 Chen EY-T, Leontyeva Y, Lin C-N, Wang J-D, Clements MS, Dickman PW. Comparing Survival Extrapolation within All-Cause and Relative Survival Frameworks by Standard Parametric Models and Flexible Parametric Spline Models Using the Swedish Cancer Registry . *Med Decis Making* 2024; published online Feb 5. DOI:doi: 10.1177/0272989X241227230.

119 Office for National Statistics. National life tables: UK (2018-2020). 2021.

120 Antonia SJ, Borghaei H, Ramalingam SS, *et al.* Four-year survival with nivolumab in patients with previously treated advanced non-small-cell lung cancer: a pooled analysis. *Lancet Oncol* 2019; **20**: 1395.

121 Vokes EE, Ready N, Felip E, *et al.* Nivolumab versus docetaxel in previously treated advanced non-small-cell lung cancer (CheckMate 017 and CheckMate 057): 3-year update and outcomes in patients with liver metastases. *Ann Oncol* 2018; **29**: 959–65.

122 Bullement A, Cranmer HL, Shields GE. A Review of Recent Decision-Analytic Models Used to Evaluate the Economic Value of Cancer Treatments. *Appl Health Econ Health Policy* 2019; **17**: 771–80.

123 Matcho A, Ryan P, Fife D, Reich C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model. *Drug Saf* 2014; **37**: 945–59.

124 Clinical Practice Research Datalink. CPRD GOLD March 2021 (Version 2021.03.001) [Data set]. Clinical Practice Research Datalink. 2021.

125 Requena G, Wolf A, Williams R, *et al.* Feasibility of using Clinical Practice Research Datalink data to identify patients with chronic obstructive pulmonary disease to enrol into real-world trials. *Pharmacoepidemiol Drug Saf* 2021; **30**: 472.

126 Ludmir EB, Mainwaring W, Lin TA, *et al.* Factors Associated With Age Disparities Among Cancer Clinical Trial Participants. *JAMA Oncol* 2019; **5**: 1769–73.

127 Zhao S, Miao M, Wang Q, Zhao H, Yang H, Wang X. The current status of clinical trials on cancer and age disparities among the most common cancer trial participants. *BMC Cancer* 2024; **24**. DOI:10.1186/S12885-023-11690-9.

128 Bullement A, Willis A, Amin A, Schlichting M, Hatswell AJ, Bharmal M. Evaluation of survival extrapolation in immuno-oncology using multiple pre-planned data cuts: Learnings to aid in model selection. *BMC Med Res Methodol* 2020; **20**. DOI:10.1186/S12874-020-00997-X.

129 Templeton AJ, Amir E, Tannock IF. Informative censoring — a neglected cause of bias in oncology trials. *Nat Rev Clin Oncol* 2020; **17**: 327–8.

130 Rosen K, Prasad V, Chen EY. Censored patients in Kaplan–Meier plots of cancer drugs: An empirical analysis of data sharing. *Eur J Cancer* 2020; **141**: 152–61.

131 Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal* 2014; **72**: 219.

132 Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester: Wiley, 2004.

133 Grant TS, Burns D, Kiff C, Lee D. A Case Study Examining the Usefulness of Cure Modelling for the Prediction of Survival Based on Data Maturity. *Pharmacoeconomics* 2020; **38**: 385–95.

134 Husereau D, Drummond M, Augustovski F, *et al.* Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) Statement: Updated Reporting Guidance for Health Economic Evaluations. *Value in Health* 2022; **25**: 3–9.

135 NHS. Understanding routes to NICE health technology assessment. 2023; published online Nov 2. https://www.digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-developers/all-developers-guidance/understanding-routes-to-nice-health-technology-assessment/ (accessed Jan 7, 2025).

136 Clinical Practice Research Datalink. Primary care data for public health research. 2024; published online Jan 23. https://cprd.com/primary-care-data-public-health-research (accessed Feb 15, 2024).

137 Clinical Practice Research Datalink. Release Notes: CPRD Aurum December 2023. 2023.

Appendices

Chapter 2 appendix

PubMed search string

(cost-effectiveness[tiab] OR ("cost effectiveness"[tiab])) AND ("external validity"[tiab] OR ("externally valid"[tiab] OR (generali*[tiab] OR ("real world"[tiab] OR ("real-world"[tiab] OR (transport*[tiab] OR (population-based[tiab] OR ("population based"[tiab])))))))) AND ("controlled trial"[tiab] OR (RCT[tiab] OR ("clinical trial"[tiab] OR ("randomised controlled trial"[tiab] OR ("randomized controlled trial"[tiab] OR (trial[tiab])))))) NOT (Clinical Trial Protocol[pt] OR (systematic review[pt] OR (protocol[tiab] OR ("systematic review"[tiab]))))

Table A.2.1. Parameters generalised per study, showing non mutually exclusive categories

AEs: adverse events, HRQoL: Health-related quality-of-life. Note references are as presented in Chapter 2.

Study	Treatment effects	Baseline risks	Safety/AEs	Resource use	HRQoL	Discontinuation or adherence	Prescribing practices
de Jong et al. (2019)	✓	✓	✓				
de Pouvourville et al. (2019)	✓	✓	✓	✓			
Yeung et al. (2017)	✓	✓					
Hartman et al. (2015)	✓	✓					
Blommestein et al. (2014)	✓	✓	✓	✓	✓		
Foglia et al. (2013)	✓						
van Gils et al. (2013)	✓	✓	✓	✓			
Farahani et al. (2006)	✓	✓	✓	✓	✓	✓	
Reed et al. (2003)	✓	✓		✓			
van Kampen et al. (2017)		✓					
Seferina et al. (2017)		✓					✓
van der Linden et al. (2015)		✓					
Wouters et al. (2013)		✓					
van Staa et al. (2009)		✓	✓	✓			
Hollier-Hann et al. (2020)				✓	✓		
Desai et al. (2019)				✓			
Wan et al. (2018)				✓			
Bellows et al. (2012)				✓		✓	
Gillet et al. (2010)				✓			
Annemans et al. (2008)				✓			
Brandes et al. (2012)						✓	
Sorenson et al. (2009)		✓				✓	
Morais et al. (2014)							✓
Chevallier et al. (2014)							✓

Chapter 3 appendix

The clinical codelists used for lung cancer are listed in the table below with the corresponding SNOMED concept ID, OMOP concept ID and concept description. Only diagnosis records alone were used to identify cancer outcome for this study. As part of the broader EHDEN work package, additional codes of interest were identified and those highlighted as irrelevant, based on feedback from clinicians with oncology expertise through an iterative process during the initial stages of analyses, were removed. Additional details used to identify the lung cancer outcome can be found in a previous study.[******]

Clinical codelists

Table A.3.1. clinical codelists for lung cancer cohort definition

SNOMED CT: systematised nomenclature of medical clinical terms

Concept Id	Concept SNOMED CT Code	Concept Description
443388	363358000	Malignant tumor of lung
258369	93880001	Primary malignant neoplasm of lung
4092216	187862007	Malignant neoplasm of upper lobe of lung
4089756	187870002	Malignant neoplasm of lower lobe of lung
4151250	269464000	Malignant neoplasm of upper lobe, bronchus or lung
4246121	93827000	Primary malignant neoplasm of hilus of lung
4089754	187866005	Malignant neoplasm of middle lobe of lung
4092217	187864008	Malignant neoplasm of middle lobe, bronchus or lung
258375	109371002	Overlapping malignant neoplasm of bronchus and lung
4092218	187865009	Malignant neoplasm of middle lobe bronchus
256646	372112000	Primary malignant neoplasm of middle lobe, bronchus or lung
261236	372135002	Primary malignant neoplasm of upper lobe, bronchus or lung
433973	93729006	Primary malignant neoplasm of bronchus of left lower lobe
762426	354701000119107	Primary malignant neoplasm of left lung
762427	354741000119109	Primary malignant neoplasm of right lung
4110587	254625005	Malignant tumor of lung parenchyma
4110589	254629004	Large cell carcinoma of lung
4110590	254631008	Giant cell carcinoma of lung
4110591	254632001	Small cell carcinoma of lung

[******] Corby G, Barclay NL, Tan EH, et al. Incidence, prevalence, and survival of lung cancer in the United Kingdom from 2000-2001: a population-based cohort study. medRxiv (2024). https://doi.org/10.1101/2024.02.26.24303406

4110705	254634000	Squamous cell carcinoma of lung
4110706	254638002	Pancoast tumor
4111807	254635004	Epithelioid hemangioendothelioma of lung
4112738	254626006	Adenocarcinoma of lung
4112739	254633006	Oat cell carcinoma of lung
4115276	254637007	Non-small cell lung cancer
4140471	427038005	Epidermal growth factor receptor negative non-small cell lung cancer
4143825	426964009	Epidermal growth factor receptor positive non-small cell lung cancer
4155293	372111007	Carcinoma of lower lobe, bronchus or lung
4157454	372110008	Primary malignant neoplasm of lower lobe, bronchus or lung
4162248	372113005	Carcinoma of middle lobe, bronchus or lung
4162252	372136001	Carcinoma of upper lobe, bronchus or lung
4196724	313353007	Squamous cell carcinoma of bronchus in left lower lobe
4196725	313355000	Squamous cell carcinoma of bronchus in right lower lobe
4197581	313354001	Squamous cell carcinoma of bronchus in left upper lobe
4197582	313356004	Squamous cell carcinoma of bronchus in right middle lobe
4197583	313357008	Squamous cell carcinoma of bronchus in right upper lobe
4208307	440173001	Nonsquamous nonsmall cell neoplasm of lung
4246027	93730001	Primary malignant neoplasm of bronchus of left upper lobe
4246126	93865007	Primary malignant neoplasm of left upper lobe of lung
4246148	93991009	Primary malignant neoplasm of right lower lobe of lung
4246804	93732009	Primary malignant neoplasm of bronchus of right middle lobe
4246805	93733004	Primary malignant neoplasm of bronchus of right upper lobe
4247727	93731002	Primary malignant neoplasm of bronchus of right lower lobe
4247832	93864006	Primary malignant neoplasm of lower lobe of left lung
4307118	423050000	Large cell carcinoma of lung, TNM stage 2
4308479	423121009	Non-small cell carcinoma of lung, TNM stage 4
4310448	423295000	Squamous cell carcinoma of lung, TNM stage 1
4310703	424132000	Non-small cell carcinoma of lung, TNM stage 1
4311501	93992002	Primary malignant neoplasm of right middle lobe of lung
4311997	422968005	Non-small cell carcinoma of lung, TNM stage 3
4312274	425376008	Squamous cell carcinoma of lung, TNM stage 4
4312567	93993007	Primary malignant neoplasm of upper lobe of right lung
4312768	424970000	Large cell carcinoma of lung, TNM stage 3
4313200	423468007	Squamous cell carcinoma of lung, TNM stage 2
4313751	423600008	Large cell carcinoma of lung, TNM stage 4
4314040	424938000	Large cell carcinoma of lung, TNM stage 1
4314172	425048006	Non-small cell carcinoma of lung, TNM stage 2
4322387	425230006	Squamous cell carcinoma of lung, TNM stage 3
36686537	12235561000119100	Large cell carcinoma of left lung
36686538	12235601000119100	Large cell carcinoma of right lung
36712707	1078881000119100	Primary adenocarcinoma of lower lobe of left lung
36712708	1078901000119100	Primary adenocarcinoma of upper lobe of left lung
36712709	1078931000119100	Primary adenocarcinoma of lower lobe of right lung

36712815	12240951000119100	Squamous cell carcinoma of left lung
36712816	12240991000119100	Squamous cell carcinoma of right lung
36712981	15956381000119100	Adenocarcinoma of right lung
36713366	683991000119103	Extensive stage primary small cell carcinoma of lung
36716426	722425009	Reactive oxygen species 1 positive non-small cell lung cancer
36716500	722528008	Primary malignant neuroendocrine neoplasm of lung
36717017	1078961000119100	Primary adenocarcinoma of upper lobe of right lung
37109576	723301009	Squamous non-small cell lung cancer
37110031	724056005	Malignant neoplasm of lower lobe of right lung
37110032	724058006	Malignant neoplasm of upper lobe of left lung
37110033	724059003	Malignant neoplasm of lower lobe of left lung
37110034	724060008	Malignant neoplasm of right upper lobe of lung
37311684	822969007	Acinar cell cystadenocarcinoma of lung
37395648	67811000119102	Primary small cell malignant neoplasm of lung, TNM stage 1
37395649	67821000119109	Primary small cell malignant neoplasm of lung, TNM stage 2
37395650	67831000119107	Primary small cell malignant neoplasm of lung, TNM stage 3
37395651	67841000119103	Primary small cell malignant neoplasm of lung, TNM stage 4
40391740	189815007	Pulmonary blastoma
40492938	448993007	Carcinoma of lung
42539251	15956341000119100	Adenocarcinoma of left lung
45766129	703228009	Non-small cell lung cancer with mutation in epidermal growth factor receptor
45766131	703230006	Non-small cell lung cancer without mutation in epidermal growth factor receptor
45768879	707403002	Primary fetal adenocarcinoma of lung
45768880	707404008	Primary mixed subtype adenocarcinoma of lung
45768881	707405009	Primary adenosquamous carcinoma of lung
45768883	707408006	Primary small cell non-keratinizing squamous cell carcinoma of lung
45768884	707409003	Primary acinar cell carcinoma of lung
45768885	707410008	Primary solid carcinoma of lung
45768886	707411007	Primary papillary adenocarcinoma of lung
45768916	707451005	Primary adenocarcinoma of lung
45768917	707452003	Primary mucinous adenocarcinoma of lung
45768918	707453008	Primary clear cell squamous cell carcinoma of lung
45768919	707454002	Primary basaloid squamous cell carcinoma of lung
45768920	707456000	Primary undifferentiated carcinoma of lung
45768921	707457009	Primary spindle cell carcinoma of lung
45768922	707458004	Primary pleomorphic carcinoma of lung
45768923	707460002	Primary pseudosarcomatous carcinoma of lung
45768927	707464006	Primary myoepithelial carcinoma of lung
45768928	707466008	Primary adenoid cystic carcinoma of lung
45768929	707467004	Primary salivary gland type carcinoma of lung
45768930	707468009	Primary mixed mucinous and non-mucinous bronchiolo-alveolar carcinoma of lung
45768931	707469001	Primary non-mucinous bronchiolo-alveolar carcinoma of lung

45768932	707470000	Primary mucinous bronchiolo-alveolar carcinoma of lung
45769034	707595001	Primary mucinous cystadenocarcinoma of lung
45769035	707596000	Primary carcinosarcoma of lung
45772933	707407001	Primary signet ring cell carcinoma of lung
45772938	707455001	Primary papillary squamous cell carcinoma of lung
45772939	707465007	Primary mucoepidermoid carcinoma of lung
46272955	711414003	Primary clear cell adenocarcinoma of lung

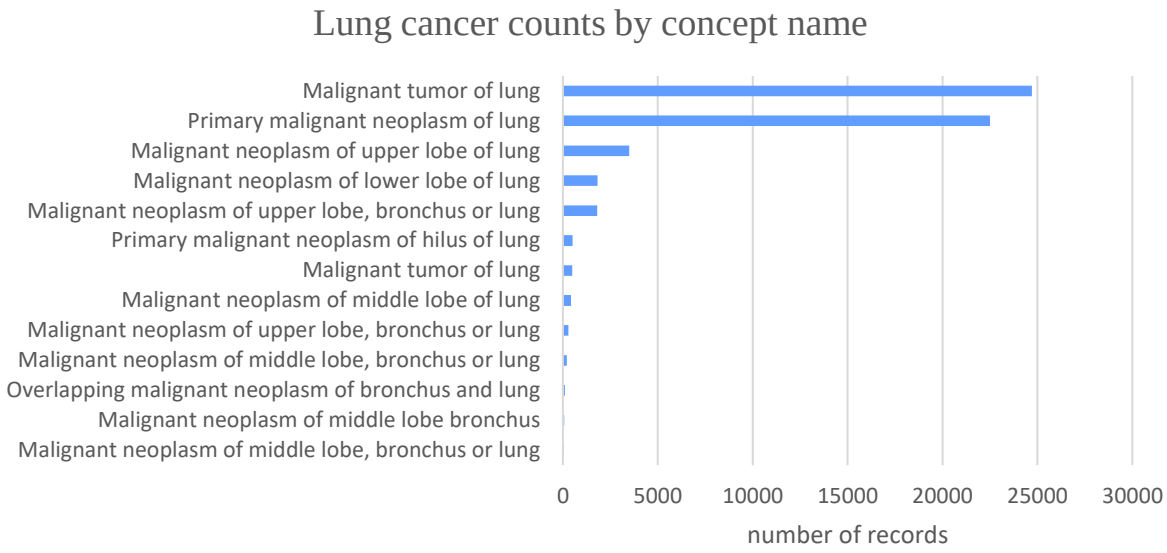


Figure A.3.1. Concepts in CPRD GOLD for lung cancer
Reproduced with permissions from Newby D. Incidence, prevalence and survival of lung cancer in the United Kingdon between 2000 and 2021. Web app companion to Corby G, Barclay NL, Tan EH, et al. Incidence, prevalence, and survival of lung cancer in the United Kingdom from 2000-2001: a population-based cohort study. medRxiv (2024). available at: https://dpa-pde-oxford.shinyapps.io/EHDENCancerIncPrevCohortDiagShiny/

Chapter 4 appendix

Sensitivity of placement of treatment effects on selective knots

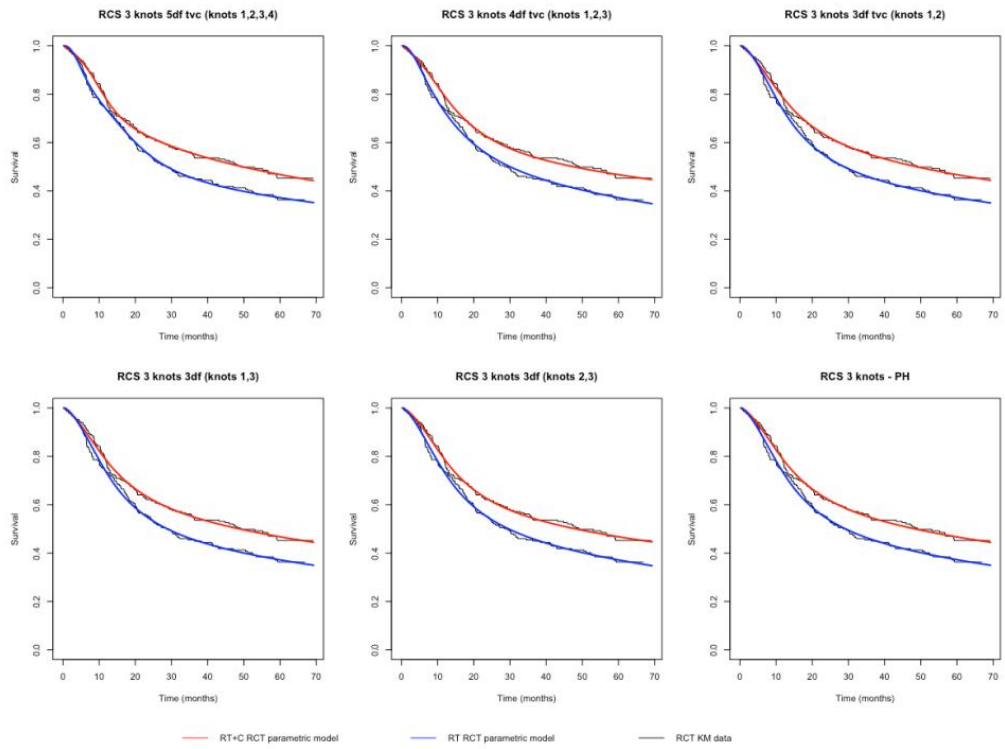


Figure A.4.1. Sensitivity to placement of relative effects on particular higher-order terms
RCS: restricted cubic spline, df: degrees of freedom

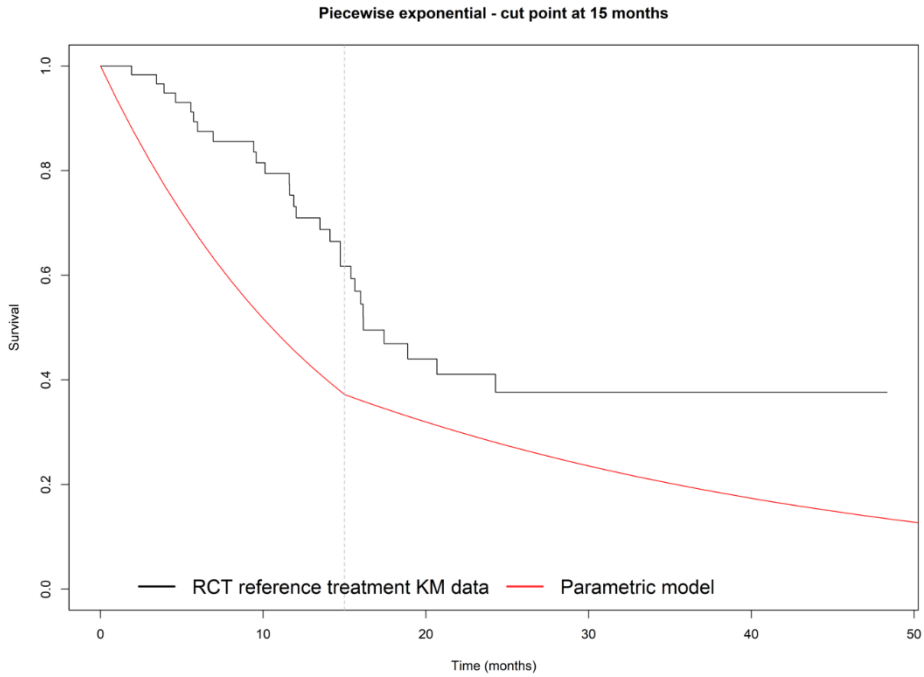


Figure A.4.2. Piecewise model fitted to baseline distribution with segments from 0-15 and >15 months

Chapter 5 appendix

Select parametric survival models fit to full CPRD lung-cancer data

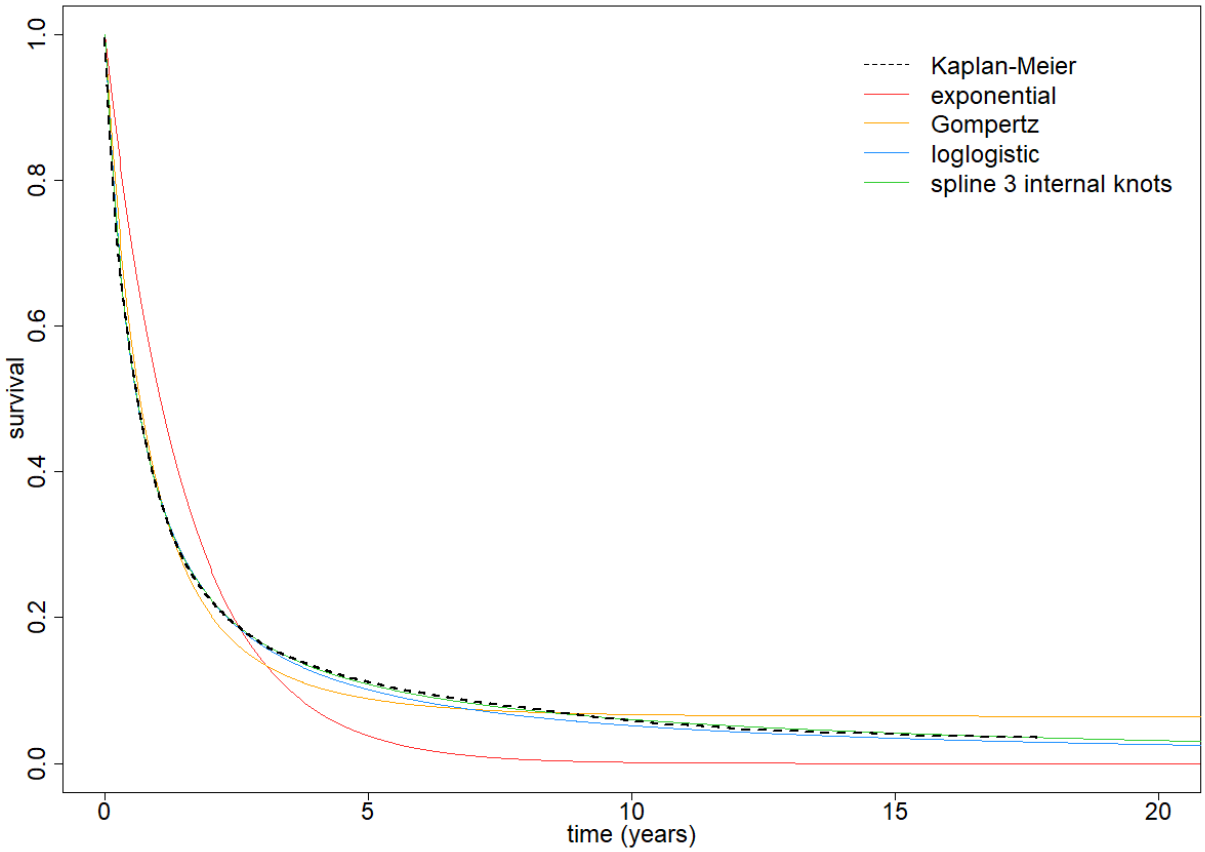


Figure A.5.1. Selected survival models fitted to full real-world dataset

Table A.5.1. Goodness-of-fit values for full real-world dataset

Survival model	AIC	BIC	AIC rank	BIC rank
exponential	331454.8	331463.3	11	11
Weibull (PH)	321794.1	321811.2	10	10
Gompertz	319329.2	319346.3	9	9
Lognormal	318122.4	318139.5	8	7
Loglogistic	318005.2	318022.3	6	6
Generalised gamma	318117	318142.7	7	8
Spline (1 internal knot)	317970.7	317996.4	5	5
Spline (2 internal knots)	317922.8	317957.0	4	4
Spline (3 internal knots)*	317807.8	317850.6	2	2
Spline (4 internal knot)	317809.6	317861.0	3	3
Spline (5 internal knot)	317783.2	317843.1	1	1

*I used a RCS3 model based on signs of overfitting with the RCS4 increasing in AIC/BIC values with no visual improvement in approximation to the observed data.

Chapter 6 appendix

Table and figures for n=150 sample size

Table A.6.1. Percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=150 sample size, for each scenario

	12 months follow-up			24 months follow-up			36 months follow-up		
Model	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring
exponential	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weibull	0.002	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
Gompertz	0.078	0.013	0.000	0.156	0.120	0.024	0.178	0.184	0.067
loglogistic	0.199	0.228	0.340	0.134	0.378	0.674	0.078	0.270	0.578
lognormal	0.091	0.499	0.566	0.043	0.143	0.148	0.031	0.097	0.128
generalised gamma	0.043	0.007	0.001	0.031	0.007	0.000	0.022	0.012	0.003
spline hazards 1 knot	0.015	0.001	0.000	0.025	0.001	0.000	0.042	0.011	0.001
spline hazards 2 knots	0.006	0.000	0.000	0.029	0.001	0.000	0.058	0.007	0.000
spline hazards 3 knots	0.017	0.000	0.000	0.030	0.000	0.000	0.072	0.009	0.000
spline normal 1 knot	0.045	0.010	0.000	0.027	0.021	0.000	0.019	0.025	0.001
spline normal 2 knots	0.030	0.002	0.000	0.042	0.003	0.000	0.028	0.007	0.000
spline normal 3 knots	0.040	0.001	0.000	0.042	0.000	0.000	0.042	0.013	0.000
spline odds 1 knot	0.217	0.230	0.091	0.186	0.294	0.153	0.179	0.290	0.206
spline odds 2 knots	0.098	0.007	0.001	0.086	0.018	0.001	0.084	0.034	0.009
spline odds 3 knots	0.119	0.002	0.001	0.168	0.014	0.000	0.166	0.041	0.007

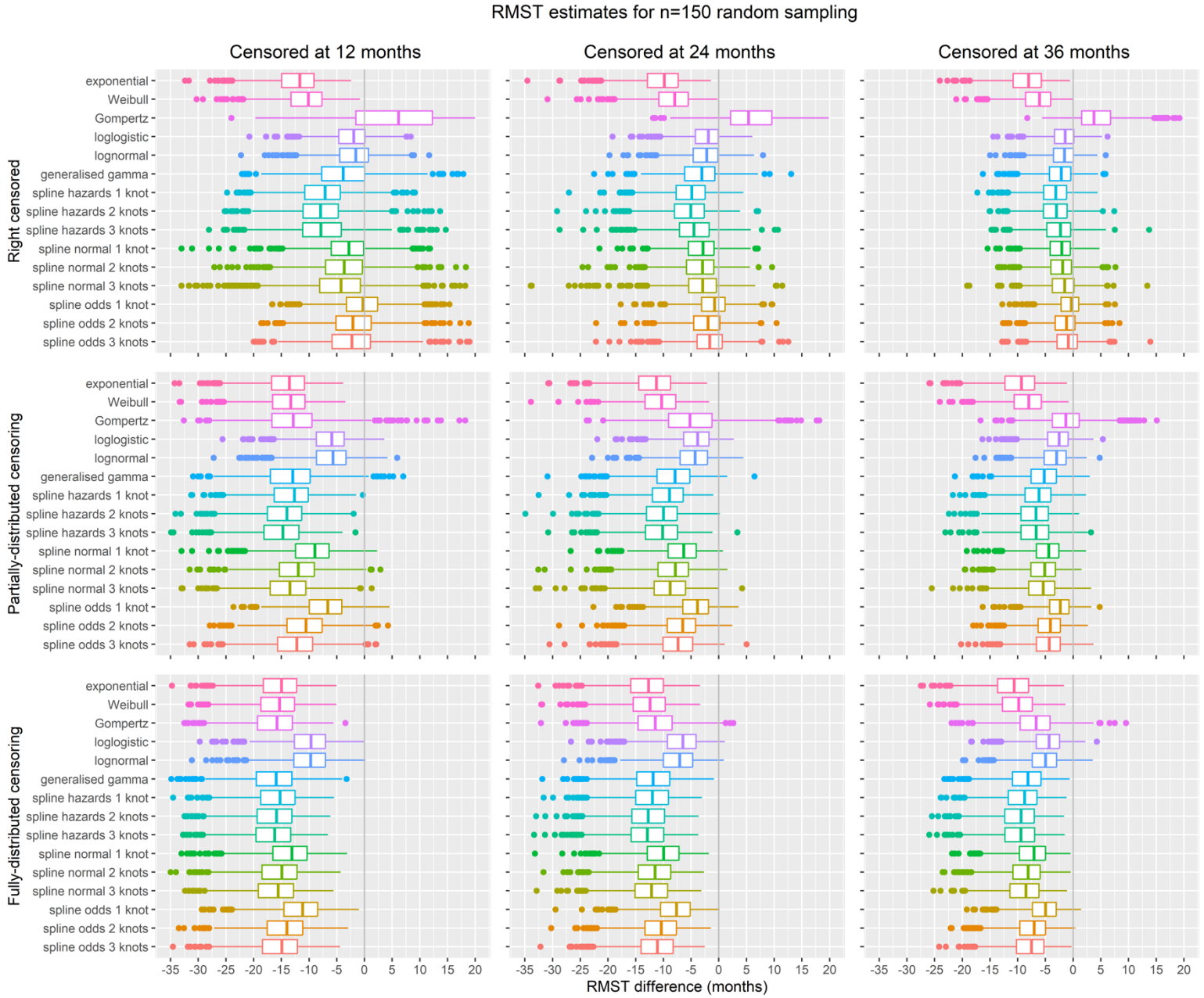


Figure A.6.1. Median, interquartile range, and range of RMST differences for the n=150 sample size, for each scenario

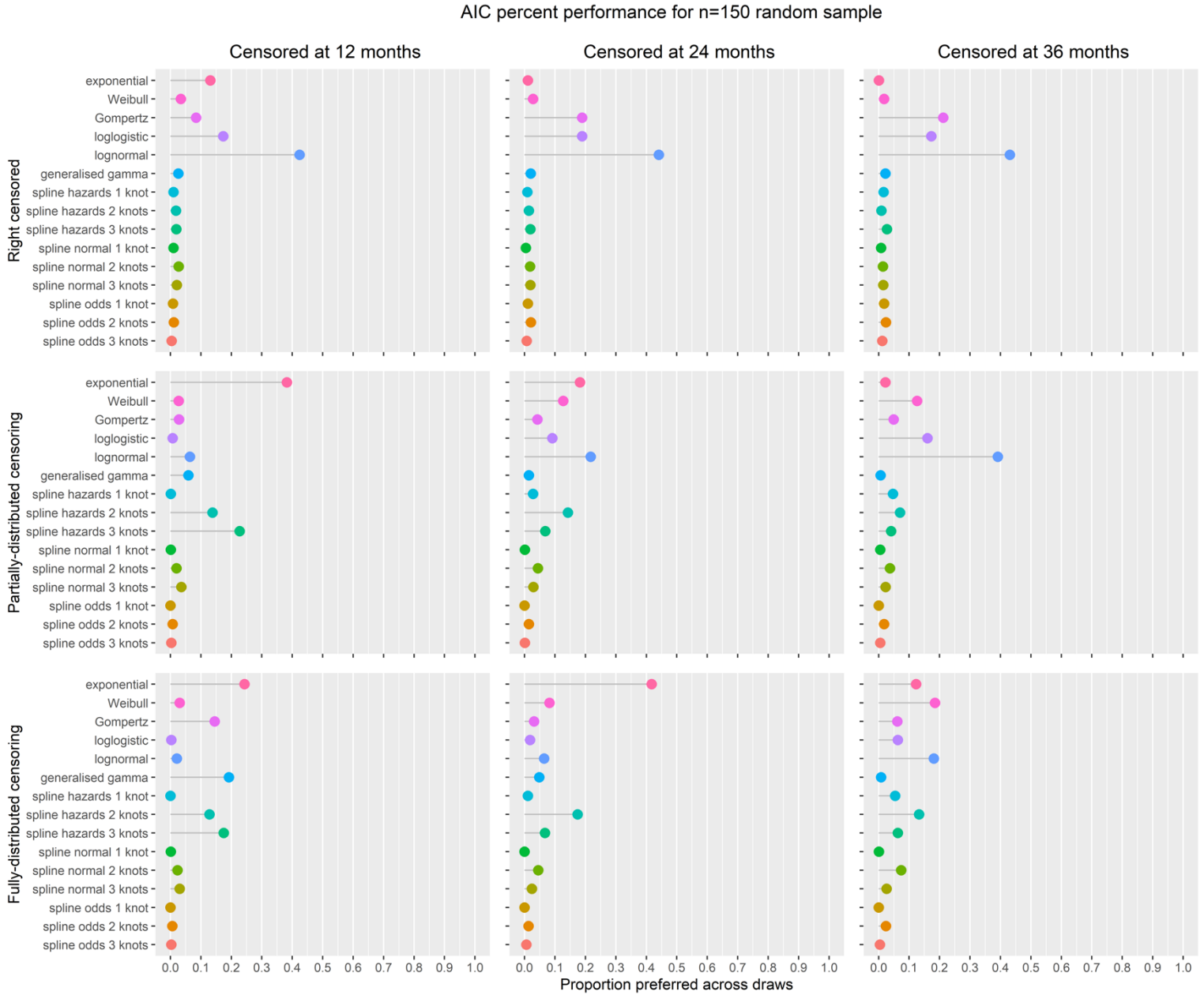


Figure A.6.2. Percentage of instances in which models exhibited the best performance, in terms of AIC, relative to the total number of draws for the n=150 sample size, for each scenario

Table and figures for n=300 sample size

Table A.6.2. Percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=300 sample size, for each scenario

	12 months follow-up			24 months follow-up			36 months follow-up		
Model	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring
exponential	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weibull	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gompertz	0.092	0.000	0.000	0.163	0.035	0.001	0.115	0.292	0.024
loglogistic	0.106	0.119	0.280	0.090	0.518	0.797	0.030	0.310	0.770
lognormal	0.171	0.728	0.703	0.046	0.196	0.154	0.013	0.079	0.082
generalised gamma	0.038	0.000	0.000	0.020	0.000	0.000	0.010	0.001	0.000
spline hazards 1 knot	0.003	0.000	0.000	0.001	0.000	0.000	0.024	0.000	0.000
spline hazards 2 knots	0.002	0.000	0.000	0.010	0.000	0.000	0.033	0.000	0.000
spline hazards 3 knots	0.000	0.000	0.000	0.033	0.000	0.000	0.107	0.003	0.000
spline normal 1 knot	0.036	0.000	0.000	0.010	0.003	0.000	0.013	0.003	0.000
spline normal 2 knots	0.005	0.000	0.000	0.011	0.000	0.000	0.023	0.002	0.000
spline normal 3 knots	0.016	0.000	0.000	0.057	0.000	0.000	0.066	0.004	0.000
spline odds 1 knot	0.385	0.153	0.017	0.294	0.247	0.048	0.191	0.273	0.117
spline odds 2 knots	0.061	0.000	0.000	0.033	0.000	0.000	0.039	0.005	0.001
spline odds 3 knots	0.085	0.000	0.000	0.232	0.001	0.000	0.336	0.028	0.006

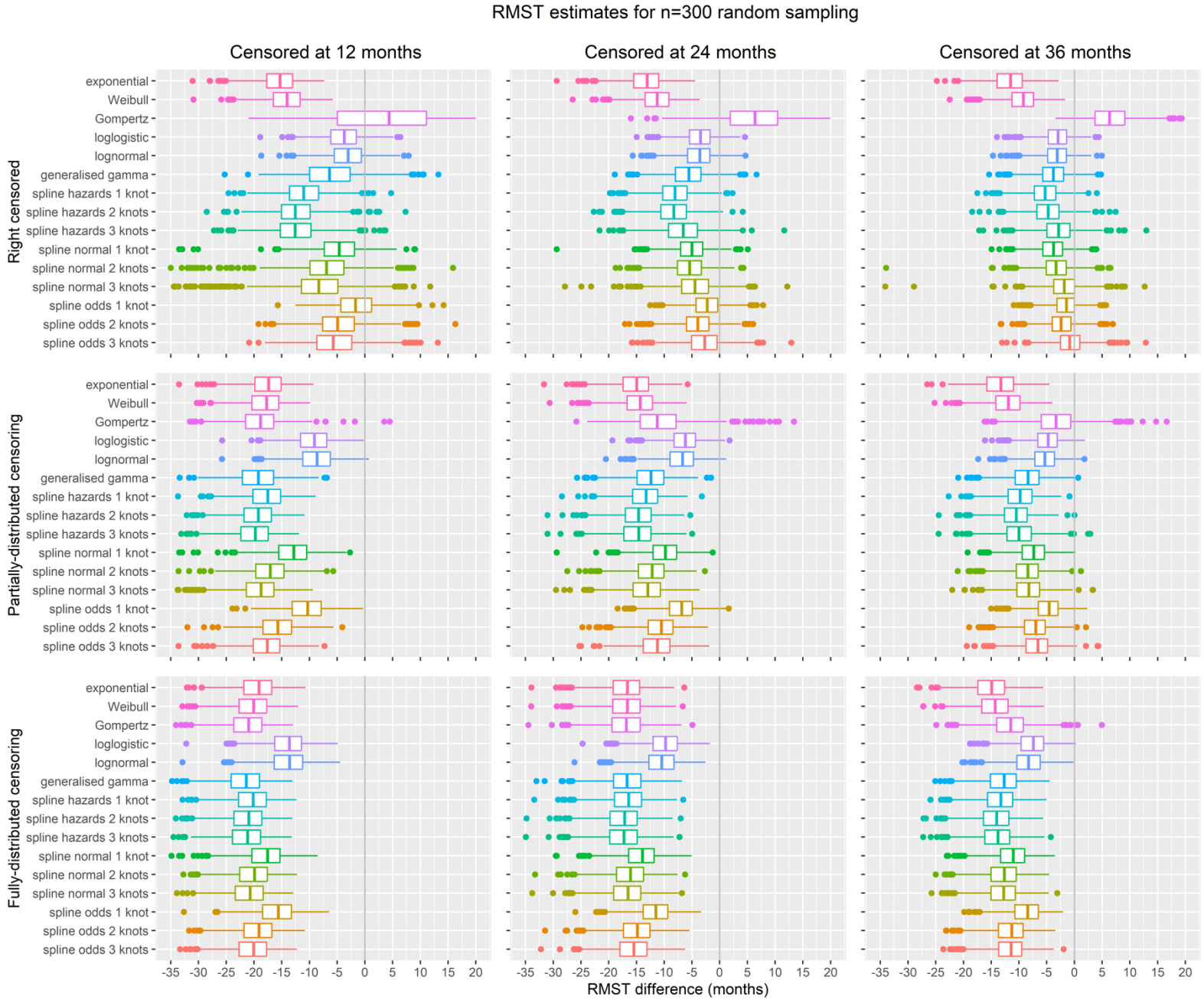


Figure A.6.3. Median, interquartile range, and range of RMST differences for the n=300 sample size, for each scenario

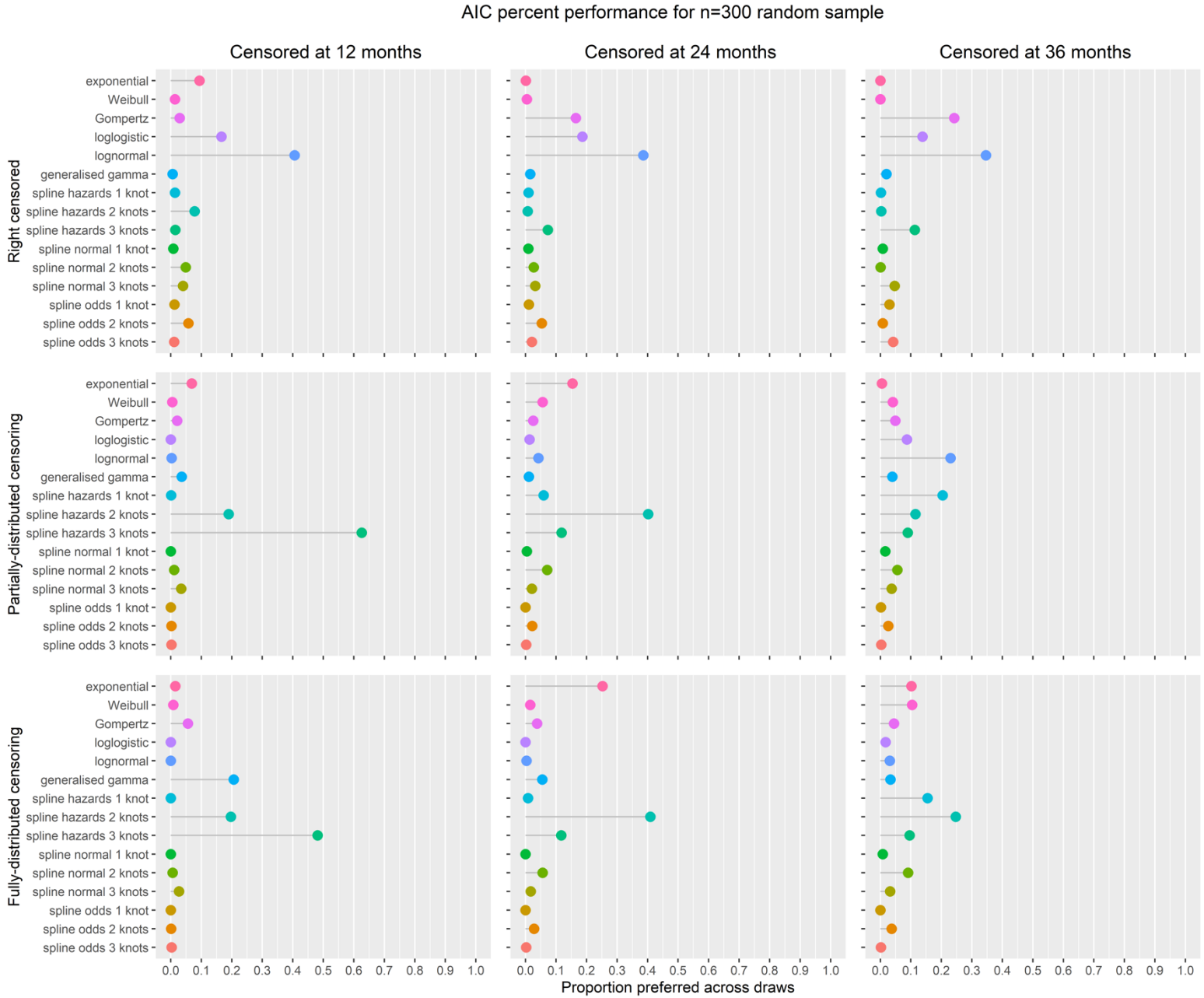


Figure A.6.4. Percentage of instances in which models exhibited the best performance, in terms of AIC, relative to the total number of draws for the n=300 sample size, for each scenario

Table and figures for n=1000 sample size

Table A.6.3. Percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the n=1000 sample size, for each scenario

	12 months follow-up			24 months follow-up			36 months follow-up		
Model	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring
exponential	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weibull	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gompertz	0.080	0.000	0.000	0.006	0.048	0.000	0.009	0.373	0.004
loglogistic	0.042	0.038	0.145	0.068	0.628	0.983	0.012	0.310	0.957
lognormal	0.099	0.918	0.855	0.040	0.028	0.006	0.005	0.016	0.003
generalised gamma	0.012	0.000	0.000	0.003	0.000	0.000	0.002	0.000	0.000
spline hazards 1 knot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
spline hazards 2 knots	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000
spline hazards 3 knots	0.000	0.000	0.000	0.002	0.000	0.000	0.071	0.000	0.000
spline normal 1 knot	0.007	0.000	0.000	0.009	0.000	0.000	0.001	0.000	0.000
spline normal 2 knots	0.001	0.000	0.000	0.011	0.000	0.000	0.007	0.000	0.000
spline normal 3 knots	0.001	0.000	0.000	0.036	0.000	0.000	0.119	0.000	0.000
spline odds 1 knot	0.705	0.044	0.000	0.607	0.296	0.011	0.213	0.299	0.036
spline odds 2 knots	0.038	0.000	0.000	0.068	0.000	0.000	0.019	0.000	0.000
spline odds 3 knots	0.015	0.000	0.000	0.150	0.000	0.000	0.540	0.002	0.000

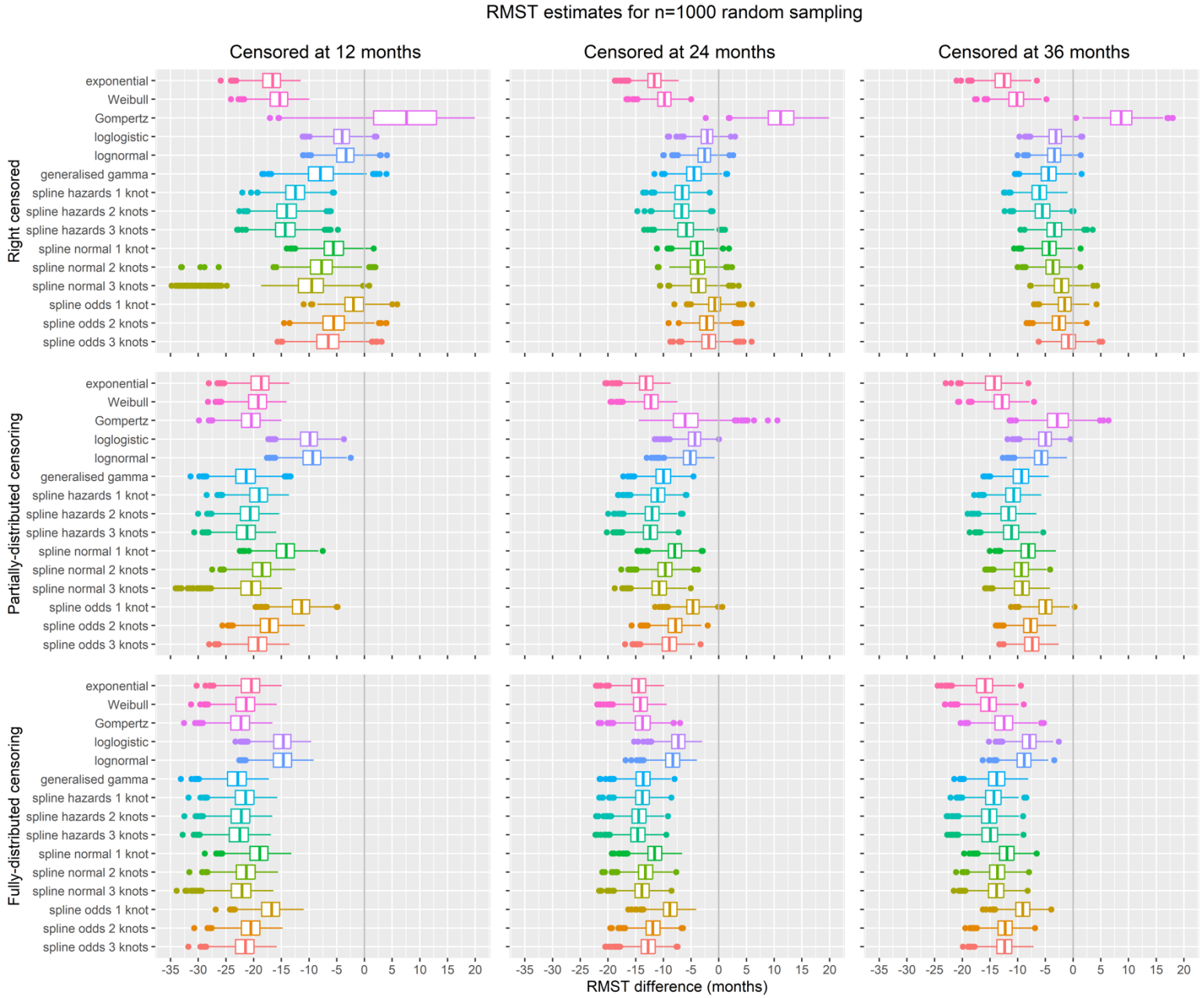


Figure A.6.5. Median, interquartile range, and range of RMST differences for the n=1000 sample size, for each scenario

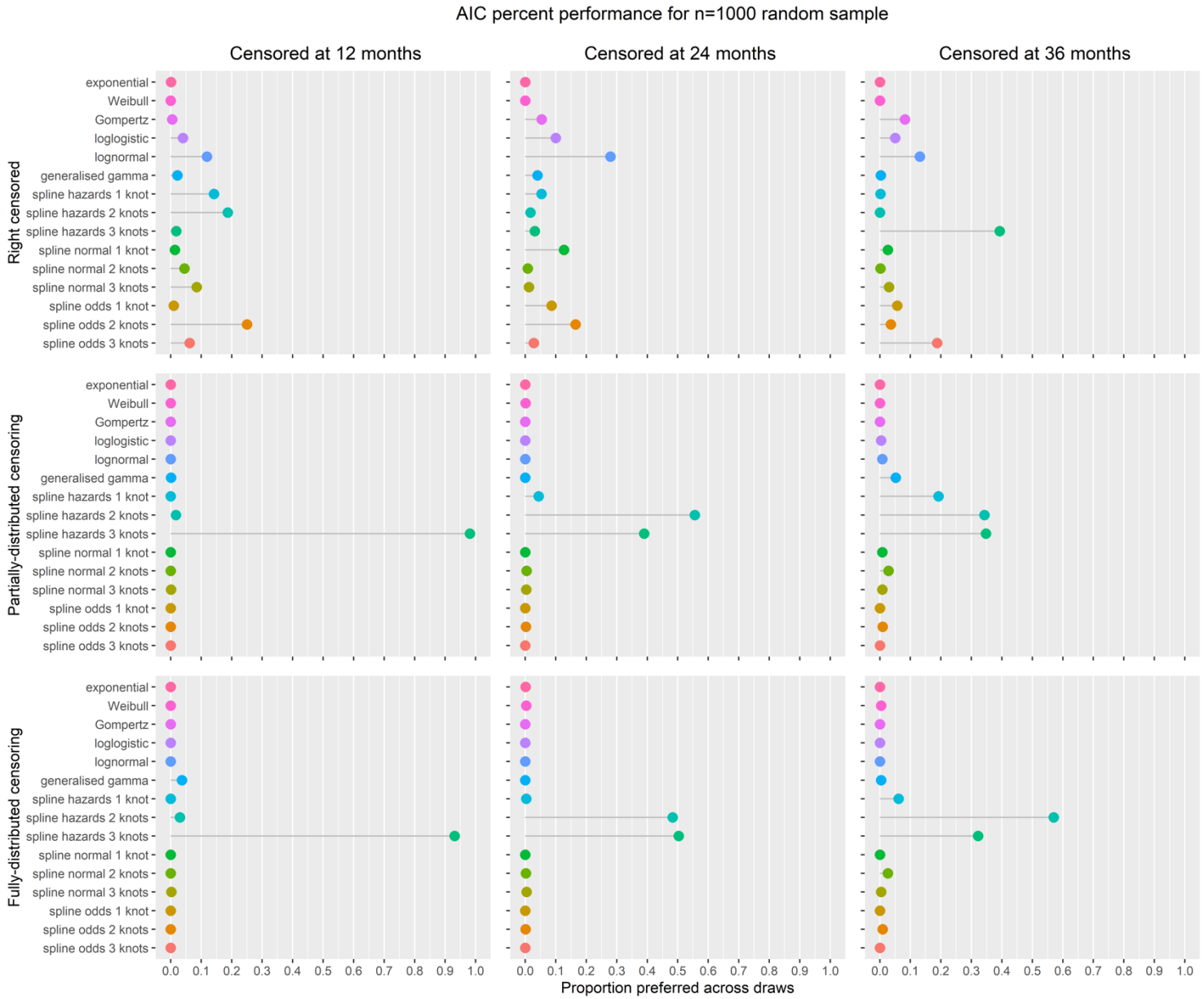


Figure A.6.6. Percentage of instances in which models exhibited the best performance, in terms of AIC, relative to the total number of draws for the n=1000 sample size, for each scenario

Table and figures for full sample size

Table A.6.4. Percentage of instances in which models exhibited the best performance, in terms of rMSE, relative to the total number of draws for the full sample size, for each scenario

	12 months follow-up			24 months follow-up			36 months follow-up		
Model	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring	Right-censored	Partially-distributed censoring	Fully-distributed censoring
exponential	0	0	0	0	0	0	0	0	0
Weibull	0	0	0	0	0	0	0	0	0
Gompertz	0	0	0	0	0	0	0	1	0
loglogistic	0	0	0	0	1	1	0	0	1
lognormal	0	1	1	0	0	0	0	0	0
generalised gamma	0	0	0	0	0	0	0	0	0
spline hazards 1 knot	0	0	0	0	0	0	0	0	0
spline hazards 2 knots	0	0	0	0	0	0	0	0	0
spline hazards 3 knots	0	0	0	0	0	0	0	0	0
spline normal 1 knot	0	0	0	0	0	0	0	0	0
spline normal 2 knots	0	0	0	0	0	0	0	0	0
spline normal 3 knots	0	0	0	0	0	0	0	0	0
spline odds 1 knot	1	0	0	1	0	0	0	0	0
spline odds 2 knots	0	0	0	0	0	0	0	0	0
spline odds 3 knots	0	0	0	0	0	0	1	0	0

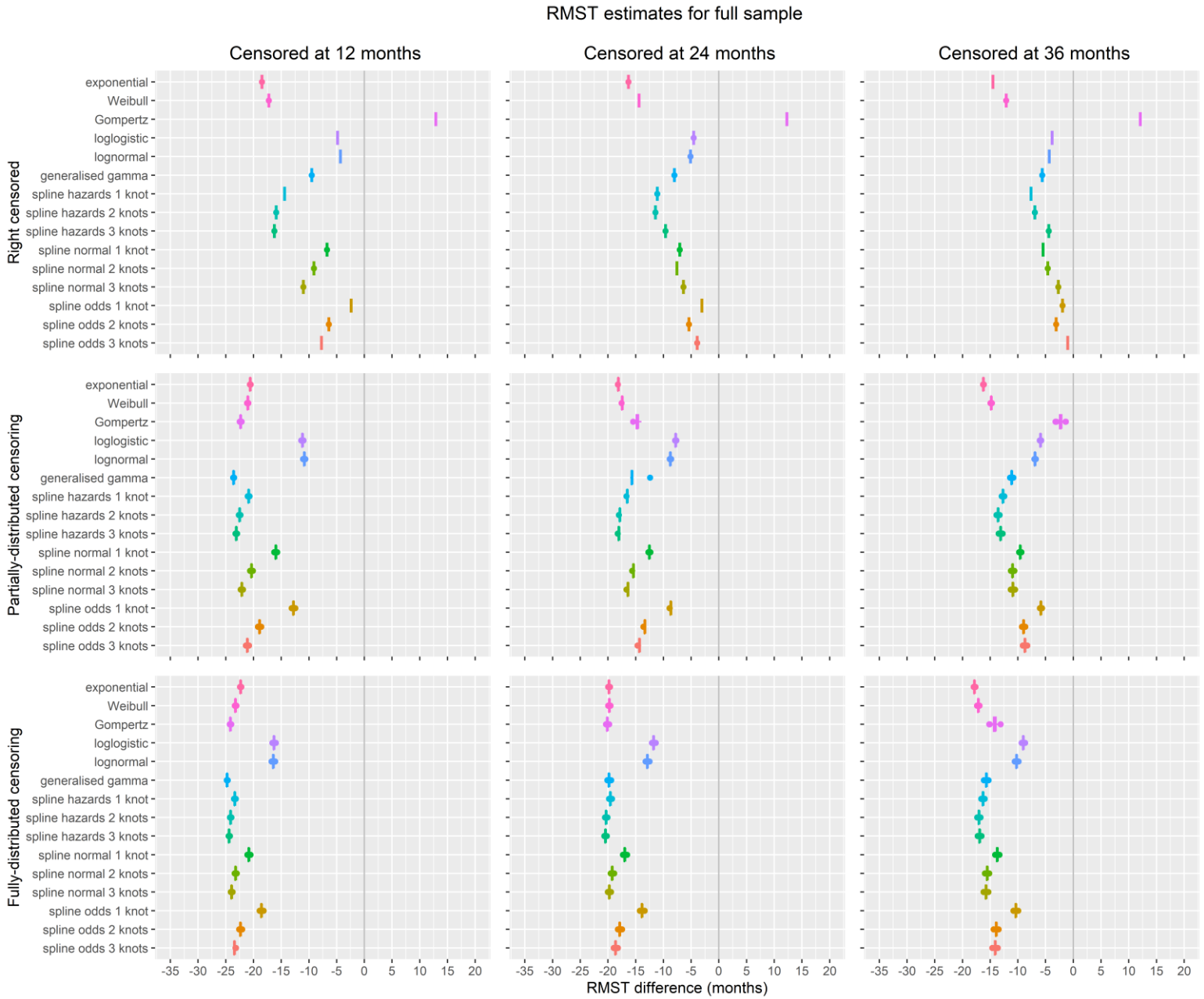


Figure A.6.7. Median, interquartile range, and range of RMST differences for the full sample size, for each scenario

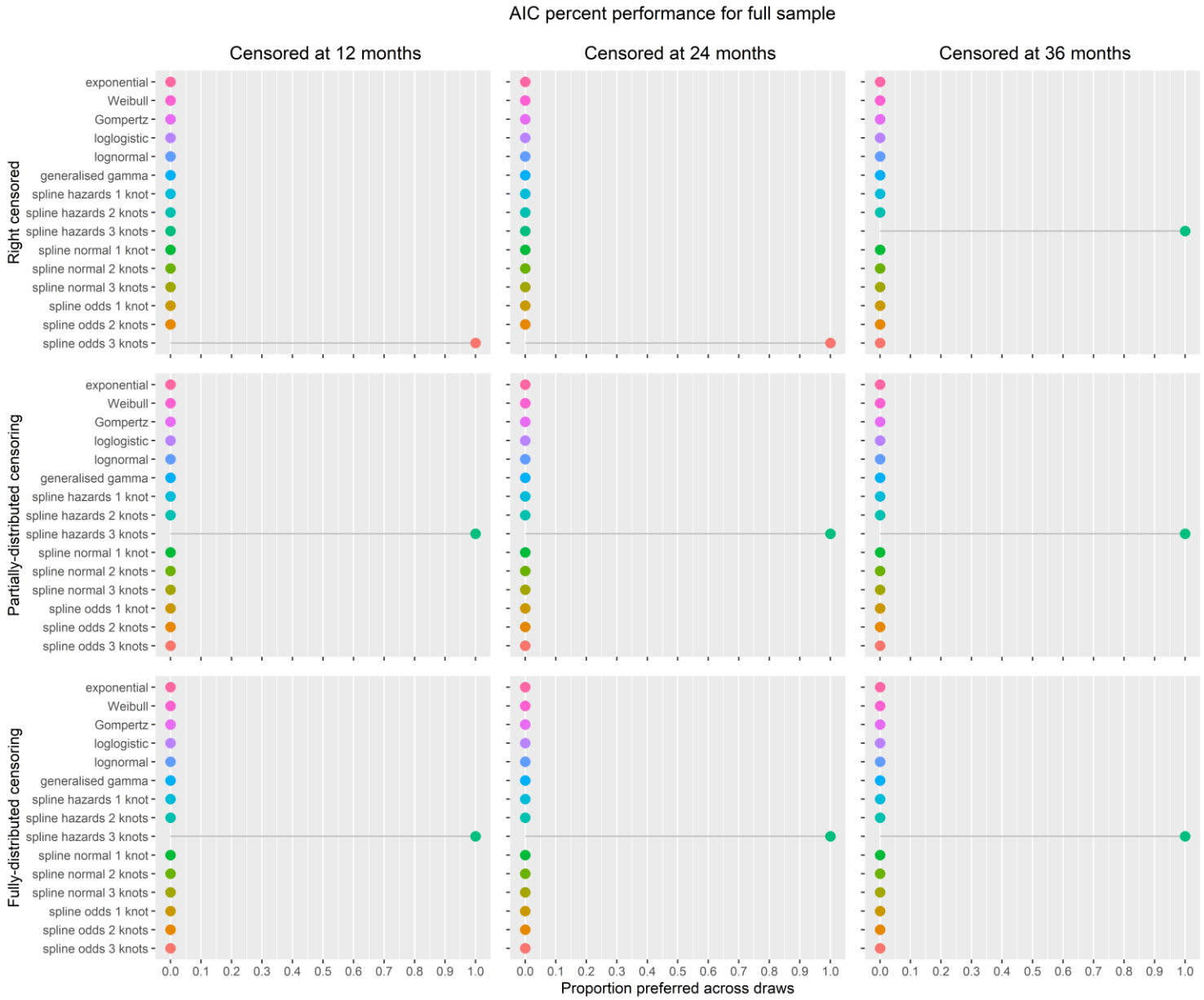


Figure A.6.8. Percentage of instances in which models exhibited the best performance, in terms of AIC, relative to the total number of draws for the full sample size, for each scenario