

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/175656>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Methods for survival extrapolation and decision making in health technology assessment

Daniel Gallacher MSc, BSc

A thesis submitted in partial fulfilment of the requirement of a Doctor of Philosophy in Health Sciences by published works.

Warwick Medical School, University of Warwick

April 2022

Table of Contents

LIST OF TABLES/FIGURES	1
ACKNOWLEDGEMENTS.....	2
DECLARATION	2
WORD COUNT.....	2
LIST OF PUBLICATIONS	3
ABSTRACT	4
1. OVERVIEW.....	5
2. BACKGROUND	7
2.1. NICE HEALTH TECHNOLOGY ASSESSMENT	7
2.2. SURVIVAL ANALYSIS AND HETEROGENEITY IN HTA.....	12
3. PUBLISHED WORKS.....	18
3.1. ESTABLISHING CURRENT PRACTICE OF SURVIVAL ANALYSIS WITHIN HEALTH TECHNOLOGY APPRAISALS – PAPER ONE.....	18
3.2. ASSESSING PRACTICE OF SURVIVAL MODELLING IN COST-EFFECTIVENESS MODELLING OUTSIDE OF TECHNOLOGY APPRAISALS – PAPER TWO	20
3.3. ASSESSING EFFICACY OF PARAMETRIC EXTRAPOLATION – PAPER THREE.....	22
3.4. CONSIDERING THE IMPACT OF ASSESSING PLAUSIBILITY AND EXPLORING MODEL AVERAGING – PAPER FOUR.....	24
3.5. THE PROBLEM OF EXTRAPOLATING FROM HETEROGENEOUS POPULATIONS – PAPER FIVE.....	26
3.6. DOES NICE’S APPRAISAL PROCESS DISCOURAGE THE DEVELOPMENT OF TARGETED THERAPIES? – PAPER SIX	28
4. DISCUSSION AND RECOMMENDATIONS	30
5. CONCLUSION	41
ACRONYMS.....	42
REFERENCES.....	43
APPENDIX A	48
APPENDIX B	54
APPENDIX C	57

List of Tables/Figures

Table 1: Overview of papers included in this thesis	3
---	---

Acknowledgements

I'm thankful for the supervision of Dr Peter Kimani and Professor Nigel Stallard whose wisdom and experience have helped increase the quality of my research and enabled me to submit this thesis. Thank you to Dr Amy Grove and Dr Paul Sutcliffe whose directorship of Warwick Evidence allowed me to experience and enjoy the work which inspired this thesis and to find the time to complete these works. I'm grateful to the support of my colleagues, co-authors and reviewers whose advice and support have helped me get this far. I'm thankful to my parents for encouraging me down the academic route and for believing I could get there. To my wife Kate, thank you for your love, patience, and encouragement over the last couple of years.

Finally, I'm grateful to God for His every blessing on my life.

Declaration

I, Daniel Gallacher, declare that:

this thesis is my own work, and that I have made clear where the contributing work has been collaborative and included signed statements indicating my contribution.

the work of this thesis has not been published beyond the six papers included in this thesis.

this thesis has not been submitted for a degree at another university.

this thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Word Count

11,613 words, excluding tables, abstract and references.

List of Publications

A list of the six papers included as part of this thesis are shown below. Details of the thesis author's contribution for each paper are included in appendix A.

Table 1: Overview of papers included in this thesis

Paper Title	Authors	Journal and Publication Date
How Do Pharmaceutical Companies Model Survival of Cancer Patients? A Review of NICE Single Technology Appraisals in 2017	D Gallacher , P Auguste, M Connock	International Journal of Technology Assessment in Health Care, 2019
A Systematic Review of Economic Evaluations Assessing the Cost-Effectiveness of Licensed Drugs Used for Previously Treated Epidermal Growth Factor Receptor (EGFR) and Anaplastic Lymphoma Kinase (ALK) Negative Advanced/Metastatic Non-Small Cell Lung Cancer	D Gallacher , P Auguste, P Royle, H Mistry, X Armoiry	Clinical Drug Investigation, 2019
Extrapolating Parametric Survival Models in Health Technology Assessment: A Simulation Study	D Gallacher , P Kimani, N Stallard	Medical Decision Making, 2021
Extrapolating Parametric Survival Models in Health Technology Assessment Using Model Averaging: A Simulation Study	D Gallacher , P Kimani, N Stallard	Medical Decision Making, 2021
Biased survival predictions when appraising health technologies in heterogeneous populations	D Gallacher , P Kimani, N Stallard	PharmacoEconomics, 2022
Development of a model to demonstrate the impact of National Institute of Health and Care Excellence cost-effectiveness assessment on health utility for targeted medicines	D Gallacher , N Stallard, P Kimani, E Gokalp, J Branke	Health Economics, 2022

Abstract

The National Institute for Health and Care Excellence (NICE) is the agency for England and Wales responsible for approving medical technologies for routine use in the National Health Service (NHS). When agencies such as NICE assess the value of health technologies, it is common to rely on predictions and extrapolations to assess their lifetime costs and benefits. The papers featured in this thesis set out to identify and assess the suitability of common extrapolation methods and explore the impact of their implementation within economic evaluations.

A systematic search identified methods of survival extrapolation used in recent NICE technology appraisals. A systematic review of cost-effectiveness studies identified survival methods used outside of NICE appraisals. Monte Carlo simulations explored utility of these methods across multiple scenarios. An economic model was built to investigate whether existing NICE processes incentivise the development of stratified therapies when there is the possibility of a heterogeneous response within a population.

Simulations demonstrated that life-year estimates obtained from routinely used parametric extrapolations were associated with bias and large imprecision. In heterogeneous populations, the bias was more severe. Averaging methods offered an improvement, generally reducing the error and bias of life-year estimates, but the variance remains high. In heterogeneous populations, stakeholders may disagree on their preference for a drug to be developed as a targeted therapy.

Current extrapolation methods are unsuitable for the major role they play in influencing healthcare decision-making. Decisions that rely on parametric extrapolations should encourage continued data collection and be regularly reviewed as new evidence becomes available. Stronger encouragement to explore subgroup effects, consistent with the recently updated NICE Methods Guide, may better incentivise the development of targeted therapies, resulting in better care for patients.

1. Overview

This thesis concerns the challenges faced when evaluating the efficacy of treatments using the methods which currently feature routinely within technology appraisals assessed by the National Institute for Health and Care Excellence (NICE). The application of statistical methodology into health technology assessment (HTA) is an emerging and evolving field. [1] Established statistical methods can be utilised in applications that they were not originally developed for, or without sufficient consideration of their suitability, and there is little evidence demonstrating whether these methods are fit for contributing to decision-making in HTA. [2]

This thesis focuses on the extrapolation of time-to-event data, where there is interest in both whether an event occurs, and when it occurs. Examples of common events include death, treatment cessation or disease progression. Survival analysis is the general term encompassing analyses of these outcomes. The decision support unit (DSU) commissioned by NICE has published technical support documents (TSDs) for performing survival analysis within economic evaluations [3, 4], but there is insufficient research to support the suggested methodology and approaches in the way that they are used in HTA.

Parametric modelling, the main method of extrapolation considered in this thesis, assumes survival times for a group of patients can be well represented by a single parametric form. Hence reliable extrapolation may be more difficult when unobserved event times do not come from the same parametric distribution. This thesis explores scenarios when parametric models are fitted to data where the survival times are sampled from either one distribution or two distinct distributions, the latter referred to as a heterogeneous population. This heterogeneity may be attributable to a combination of prognostic and treatment-effect modifying variables. Where treatment efficacy varies among different groups of patients, the therapy could be given only to those it is most beneficial to, meaning it would be described as a stratified therapy.

The 6 publications included in this thesis identify current practice in technology appraisals and also in cost-effectiveness studies published in peer-reviewed journals. They explore the suitability of these approaches through simulations, and

present for the first time in this context the method of averaging across all plausible models, rather than relying on a single extrapolation. Potential problems of survival modelling of heterogeneous populations are explored. Finally, an application of heterogeneous population modelling is explored through a representation of NICE's framework for appraising therapies. Looking at scenarios where patient response to treatment is variable, NICE's valuation process is modelled to examine whether or not it encourages the development of stratified therapies by pharmaceutical companies.

The papers in this thesis together address the research question: What is the utility and implications of current methods for extrapolating survival outcomes in NICE technology appraisals?

The objectives of the first paper were to systematically identify the methods used to model survival data in the economic analyses, establish whether the approaches adhered to the NICE DSU TSD 14, and how the submissions explored model uncertainty in the survival extrapolations. [5] The second paper was a systematic review of cost-effectiveness analyses for non-small-cell lung cancer (NSCLC), with a focus on the methods of extrapolating progression-free survival (PFS) and overall survival (OS) to investigate the methodology used beyond the setting of NICE technology appraisals. [6]

Having established current practice across both technology appraisals and published literature, the papers three and four assess its utility. [7, 8] Paper three demonstrated the impact of relying on Akaike information criterion (AIC) and Bayesian information criterion (BIC) alone, representing occasions when all candidate models could be considered plausible. The results suggested that typical phase III trial follow-up does not contain enough information to provide a reliable extrapolation, with the variance of life-year estimates increasing when a model is selected based on AIC or BIC. [7] Paper four investigated whether any improvements could be made over the approaches considered in the third paper. [8] The fourth paper introduced a range of model averaging approaches which provided an alternative to the selection of a single parametric model and considered model plausibility in combination with goodness-of-fit.

The fifth paper used parametric models and methods of averaging to extrapolate from heterogeneous populations, where heterogeneity was present either in the form of prognostic or treatment effect-modifying factors. [9] It explored the impact of increasing a trial's sample size and fitting separate survival models to subgroups when statistically significant heterogeneity was detected within a population.

Paper six investigated if NICE's valuation process discourages pharmaceutical companies from identifying heterogeneous treatment effects, beyond the bias in the survival modelling already identified.⁷ [10]

The findings and methods introduced across these papers will influence current practice, create greater awareness of the issues related in extrapolating survival models and the associated uncertainty, and improve reliability in the assessment of health technologies.

2. Background

2.1. NICE health technology assessment

In England and Wales, before a patient can access a particular treatment on the National Health Service (NHS), the treatment must be proven to be safe, be proven to be efficacious, and be deemed cost-effective by NICE. NICE is a public body responsible for improving health and social care outcomes. It achieves this by producing guidance and advice for health and care practitioners, developing standards and metrics for providers of health and care services and providing information services for those involved with health and social care delivery. One of the key forms of guidance are technology appraisals that recommend which new treatments should be made available on the NHS. Before a treatment is approved and made accessible, a submission of evidence is provided to NICE by the pharmaceutical company, which NICE pass on to one of their external assessment groups (EAGs) who provide an independent critique on the submission. EAGs are linked with academic institutions and comprise diverse teams including systematic reviewers, health economists and statisticians who investigate whether there is

evidence to support the benefit of a health technology as claimed by the company. [11] The EAG produce their own report, which is considered alongside the company submission by a committee of members from the NHS, academia, other companies and patient/carer representatives who decide whether or not the treatment is cost-effective and suitable for reimbursement on the NHS. [12]

Pharmaceutical companies are a common source of novel therapies [13] and so the responsibility to demonstrate a new technology's safety and efficacy lies with them, rather than a public government or healthcare authority. A company typically identifies multiple potential treatments for initial investigation and selects the most promising for further development and testing. After numerous rounds of sifting and preclinical testing, the best performing products are tested for safety on humans in phase I clinical trials. [14] If there are no safety concerns, patients are given the treatment as part of a phase II clinical trial to identify the most appropriate dosing regimen to achieve a positive response. Once the optimal dose has been established, and if it is ethical to do so, a phase III trial will be conducted where participants are usually randomised to either the novel intervention or an existing comparator treatment, where the trial will contain sufficient patients in order to demonstrate statistical superiority or non-inferiority of the novel therapy relative to current standard treatment on one or multiple key clinical outcomes. [15]

The cost of developing a successful therapy is undoubtedly high, with recent estimates ranging from \$985 million [16] to \$2.8 billion [17], and it is understandable that companies can only operate if they are able to not only recoup their investment in the successful therapy, but also that associated with the compounds that were unsuccessful.

To improve standards of healthcare, and incentivise the development of new therapies, NICE is willing to consider a higher associated cost for a novel treatment if that treatment offers additional benefit to existing care. [18] Unfortunately, the budgets of healthcare providers are limited, and so NICE must consider the cost-effectiveness of a new therapy in addition to its safety and efficacy. It is unlikely that healthcare providers have significant reserves of spare money earmarked for

new therapies, and so by agreeing to fund a new therapy they may be putting additional pressure onto the healthcare system, and potentially displacing other interventions.

Companies demonstrate cost effectiveness using an economic model which is included in their submission to NICE. [19] As recommended by NICE for treatments that affect costs and outcomes across a patient's lifetime, the model will usually capture all associated costs borne by the healthcare provider in caring for patients from the point of beginning the therapy until all patients are expected to have died. Depending on the disease and patient population, this time horizon could be very short, or could span many decades if patients have good survival prospects. The clinical benefits of the treatment across the patient population are also captured in the model and are measured in quality adjusted life-years (QALYs). [19] QALYs allow for the comparison of benefit to be made across different diseases and health problems, with one QALY representing a year of perfect health and death being represented by zero QALYs. Utility values are generated using tools such as EuroQol-5D (EQ-5D) [20] which assess population preferences for different health conditions. EQ-5D measures the impact across five domains (mobility, self-care, usual activity, pain/discomfort and anxiety/depression). Some severe health states can even be given a negative score meaning they are considered worse than death. [21] The value of different health-states can also vary by country, allowing for cultural differences around the world. [22]

The main health states of a disease are represented within the economic model: for stage 3/4 cancer these are usually progression-free survival, post-progression survival and death. [23] A utility score for the quality of health expected in each health state is calculated and multiplied by the average time spent in each health state, life-years, in order to estimate the QALYs.

If survival data are sufficiently mature, then the life-years may be estimated through calculating the area under the Kaplan-Meier curve. [24] A Kaplan-Meier curve estimates the proportion of a population who are event free over time, accounting for people who may be censored and their status being unknown beyond a certain point.[25] This is introduced in more detail in Appendix C.

However, as is often the case, there will be a proportion of patients who remain alive at the end of the trial follow-up. Whilst waiting for extended follow-up from a clinical trial could be an option, healthcare providers are often under significant pressure from the media, patients and politicians to ensure rapid access to the latest treatments meaning an earlier decision is usually required. Since NICE stipulates that the benefit must be estimated across the patient lifetime, it is common for companies to predict the survival of patients beyond the observed period of trial follow-up. [24]

If a phase III trial has been conducted with a relevant comparator, the survival from both arms will be extrapolated. Sometimes a phase III trial is not feasible, or the comparator arm of the trial might be outdated and not representative of current practice. In this case, an estimate of efficacy for standard care must be obtained using other means, such as methods of indirect comparison or routine data. A reference point of existing care is necessary to estimate the incremental benefit of the new technology, which is measured in QALYs.

From the author's experience of appraising submissions received by NICE, it was observed that companies use a range of survival analysis techniques to model survival data and extrapolate into the future. These extrapolations are then used to estimate the life-years by calculating the area under the curve. However, these techniques were not developed with this application in mind, and their suitability for this use in appraising health technologies is unknown. Parametric survival models, like other explanatory models, are fitted to observed data to allow exploration of the effects of different variables and their extrapolations may misrepresent the true future behaviour. It is these observations that motivated the ideas behind the papers comprising this thesis. Where usual parametric models represent the data poorly, the population may be heterogeneous. In such a case, patients within a heterogeneous population can have varying prognoses or responses to treatment. Patients' responses or prognoses may be predicted if these outcomes can be matched to a biomarker, or biological characteristic. On some occasions, a treatment may be less effective or ineffective for patients who do not have a certain biomarker. Treatments given only to biomarker positive patients are

known as stratified therapies. The development of a stratified therapy may be more expensive due to the costs of discovery and detection of a biomarker. [26]

Regardless of whether a treatment is stratified, the benefit it provides is assessed using life-years and QALYs in the same way.

Life-years not only contribute to the estimate of treatment benefit received by patients but also affect the costs associated with each treatment, as many patient costs are associated with how long they remain alive. The additional QALYs a new treatment provides compared to existing care is known as the incremental benefit. Similarly, the incremental cost is the difference between the total costs of the intervention and comparator and, like the benefits, is estimated for the patient population across their lifetime.

The incremental cost-effectiveness ratio (ICER) is defined as the incremental costs divided by the incremental benefits. NICE assesses cost-effectiveness of most therapies using a willingness-to-pay threshold of £20,000-£30,000/QALY gained, presuming a new treatment provides more QALYs at an additional cost compared to existing care, with the new treatment deemed cost-effective if the corresponding ICER is below the threshold. [27] Special consideration is given to treatments deemed end of life therapies, where the threshold is increased to £50,000/QALY gained, and to treatments for rare diseases, referred to as highly specialised technologies. [28, 29]

Survival modelling is a fundamental tool used in the assessment of health technologies under the current approach to appraisal by NICE, [5, 24] and it is of the utmost importance that decision-makers are aware of the strengths and limitations of current methods to ensure that fair pricing and access to therapies is maintained. The performance of survival models and other statistical methods can be assessed through simulation. [30] Ideally, every statistical method will be unbiased and precise. The presence of bias means there is a systematic difference between the estimates and the truth that is skewing the results. A precise estimator for treatment effect will give similar results each time, with little variance in the estimates. Bias is assessed by comparing the mean value estimated by the estimator within the simulations to the true value. Mean-squared error assesses

the performance of an estimator examining both bias and variance of the estimator. [31]

2.2. Survival analysis and heterogeneity in HTA

Extrapolation of survival data from a clinical trial is common in HTA [24]. The survival data will often come from a trial that was designed to demonstrate treatment safety or efficacy in terms of a hazard ratio. [5, 24] The point at which the main analysis of the trial data is conducted is usually driven by the number of observed events and will not usually require every patient to have experienced the event of interest. The NICE DSU has produced a series of TSDs to provide guidance on how to implement appropriate methods for appraising health technologies. Some of these are specific to survival modelling, the most relevant being TSD 14, which outlines methods for extrapolation of patient level data, and TSD 21 which describes flexible methods of survival analysis. [3, 4]

TSD 14 outlines suitable modelling approaches, focusing largely on parametric models, but also mentioning piecewise approaches. It suggests using visual inspection of survival and log-cumulative hazard plots alongside AIC and BIC to identify the best fitting models. TSD 14 recommends that the clinical plausibility of each model should also be assessed by comparison to external data and expert clinical opinion. Importantly, it also states that it is not necessary to implement a proportional hazards assumption when modelling with patient-level data. For this reason, the implementation of the proportionality assumption is not explored in this thesis.

TSD 21 describes more advanced techniques that can be useful when the expected hazard rate is not well represented by a common parametric model. These techniques include restricted cubic splines, mixture models and cure modelling, which offer a range of alternatives to regular parametric models. TSD 21 demonstrates the efficacy of these techniques across a number of simulated scenarios.

Other more specialised documents also exist, such as TSD 16 which considers adjustments to survival time estimates to be applied when patients have switched treatments. [32]

From the author's experience, at the point of the first submission to NICE the data that are extrapolated will usually come from the main trial analysis. Occasionally the data will come from a pre-planned interim analysis if a trial is demonstrating superior efficacy early on. If NICE do not initially recommend a technology, then a company may submit additional information such as extended trial follow-up or with a new commercial agreement to try and reduce some of the uncertainties in the initial review. However, it is likely some extrapolation will always be necessary as there will be patients censored due to the limited follow-up who have not yet had the event of interest.

Extrapolation is usually performed through the fitting of parametric models to the observed time-to-event data, though other approaches are possible [33, 34]. Common models considered are the exponential, Weibull, log-logistic, log-normal, generalised gamma and Gompertz, which are described in more detail in Appendix C. [24] These each assume different parameterisations of the survival function, allowing the modelling of constant, decreasing, increasing or other variable hazard rates. The company and EAG will usually both select a preferred model from those available, basing their decision on the fit to the observed data and on the plausibility of the extrapolation. [3]

The assessment of the fit can be done visually, comparing the smooth line of the fitted model to the line of the Kaplan-Meier plot. [3] This can be subjective, and can lead to a focus on the fit in the tail of the Kaplan-Meier plot, where there are small numbers of patients at risk, and the uncertainty associated the population's survival is much higher than in the early stages of follow-up. [35] Focusing on the tail region where data are sparse and variability is high could result in selecting an extrapolation model that overfits to the tail data, and is not representative of the population's survival. [33]

An alternative approach to assess model fit is to consider the statistical goodness of fit. The two most common ways of doing this are using Akaike's information criterion (AIC) and Bayes information criterion (BIC).

The AIC of a model is defined as:

$$\text{AIC} = -2 \log(L) + 2k,$$

where L is the likelihood associated with the model, and k is the number of parameters. [36]

Similarly, the BIC of a model is defined as:

$$\text{BIC} = -2 \log(L) + 2k \log(n),$$

where n is usually defined as the number of patients in the analysis but can also be the number of observed events. [37, 38]

Both information criteria attempt to balance closeness to the observed data against the potential for overfitting, through the inclusion of the likelihood term and a penalty term containing k . Whilst the values from the two criteria should not be compared, a lower value for either indicates a superior model. Burnham and Anderson [39] suggest definitions for how to interpret differences in AIC, and Raftery [40] suggests similar interpretations for BIC.

Ideally, the visual fit of a model will be assessed first, to ensure the models are somewhat representative of the data. Once one or more visually plausible models are identified, their AIC and/or BIC values are then compared to provide further distinction.

If few or no models resemble the data, then alternative models could be considered. Flexible parametric models such as restricted cubic splines or fractional polynomials can be used to capture complex survival data, or a piecewise approach can also be considered, where the usual parametric models are fitted only to data that occurs beyond a certain point of follow-up where the model is better suited to capturing and representing the behaviour of the data. [24]

In addition to their fit to the observed data, models are also assessed on their clinical plausibility. If external data sources exist, such as routine data or earlier stage clinical trials, the follow-up from these can be used to compare against model predictions. If no sources exist, then expert clinical opinion is relied upon. This can be very subjective and unreliable, with experts disagreeing or being unable to predict the future efficacy of a novel therapy, hence the fit to observed data is often given a high weighting.

Usually, the effect of background mortality is applied to the survival extrapolations, which will ensure that implausibly optimistic extrapolations are curtailed to some extent.

AIC and BIC can be used regardless of the models being compared to identify a well-fitting parsimonious model. However, neither AIC nor BIC were specifically developed to be applied to selecting the most appropriate survival extrapolation. [36, 38] Typically, overfitting is a concern when there are a number of potential covariates to include in the model and including additional terms will increase the flexibility of the model and allow it to better fit to the data. AIC and BIC ensure that only covariates whose benefit outweighs the penalty for including an additional term are included in the final model. However, in the problem of selecting a model for the extrapolation of survival data, there are usually no covariates included in the models, with the only parameters in the model coming from underlying distributions and treatment effects.

The exponential distribution is the simplest distribution with one parameter, whilst Weibull, log-normal, log-logistic, gamma and Gompertz models each have two parameters. The generalised gamma and generalised F distributions have the most parameters having three and four respectively. Hence the problem of overfitting may not be very relevant in this setting, where the range of potential parameters in the model is small.

Usually, a single parametric model will be used to extrapolate survival data. There may be times where the observed survival is not well-represented by a single model. This might be attributable to different groups of patients having different

baseline prognoses or different responses to treatment. This is described as a heterogeneous population. When heterogeneity is present, companies can take different approaches. In the NICE appraisal of pertuzumab for adjuvant treatment of HER2-positive early breast cancer (TA569), the company initially sought approval for patients with either node-positive disease or hormone receptor negative disease. The company modelled these groups separately, however as these biomarkers are not mutually exclusive, some patients were contributing to both groups and both analyses. Later in the appraisal, the company abandoned the hormone-receptor negative subgroup and focused solely on the node-positive subgroup which received a positive recommendation.

In the NICE appraisal of pembrolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum containing chemotherapy (TA519), the company only considered a single, pooled population. This meant the population included patients with and without PD-L1 positive disease, defined as a tumour PD-L1 combined positive score above 1%, despite the hazard ratio showing association with PD-L1 status. [41] This association is consistent with pembrolizumab's mechanism of action that targets the PD-1 receptor, and the fact pembrolizumab is licensed only for patients with PD-L1 positive disease for other cancers, such as non-small-cell lung cancer. [42, 43] The cost-effectiveness of different subgroups was not explored, and pembrolizumab was eventually not recommended for any patients after spending some time in the Cancer Drugs Fund (CDF) to collect more data.

Similarly, the NICE appraisal of lisocabtagene-maraleucel for treating relapsed or refractory aggressive B-cell non-Hodgkin lymphoma (TA10477) included patients with diffuse large B-cell lymphoma (DLBCL), primary mediastinal B-cell lymphoma (PMBCL) and follicular lymphoma grade 3B (FL3B). DLBCL affects older patients, whilst PMBCL occurs in younger individuals. FL3B is rarer but is characterised as a faster growing disease than DLBCL and PMBCL. Whilst each group is treated similarly, the long-term outcomes for cured patients are not necessarily equal due to natural life expectancy, with a patient cured of PMBCL potentially gaining many more life-years relative to the other two groups. However, this appraisal only

considered cost-effective analyses for one combined patient population and did not explore subgroups. The outcome is not known.

In contrast, in the NICE appraisal of venetoclax and obinutuzumab for untreated chronic lymphocytic leukaemia (TA663), the company separately modelled two mutually exclusive subgroups based on whether or not a patients' disease had either 17p deletion or TP53 mutation, with the presence of these features being associated with a worse prognosis. Venetoclax was recommended by the NICE committee for patients with either the deletion or mutation, and was permitted for use in the CDF for patients without the features until more evidence is available.

It is unclear why there might be such a variety of approaches taken. Paget et al. present a list of "good statistical principles" for subgroup analyses in HTA but focus on the clinical effectiveness perspective. [44] The NICE DSU has published TSD 3 on heterogeneity, however this focuses on heterogeneity between trials in evidence synthesis using meta-analysis and meta-regression techniques, rather than within a trial. [45] The recently published NICE methods guide recommends exploration of subgroups where the level of treatment benefit may vary. [46] The subgroups should be based on an expectation of varying clinical- or cost-effectiveness, ideally identified early in the appraisal process. The guide mentions these subgroups can be based on differences in "baseline risk of specific health outcomes" but does not give examples. It also suggests considering using an established checklist such as by Sun et al., however these checklists are targeted at clinical effect modifiers and may miss prognostic factors that go on to become effect-modifying from a cost-effectiveness perspective. [47] It is currently unclear what effect the latest methods guide will have, but it is apparent that prior to the publication of the updated guide there were a wide variety of approaches to modelling heterogeneity. It is not known why one the company submission for one intervention may pool patients together whilst another selects a subgroup to focus on. Survival modelling or even NICE's methods of appraisal may influence the choices of pharmaceutical companies, but research is needed to better understand these possibilities.

The work described by the papers in this thesis has the following aims:

- To establish current methods used to extrapolate survival data in HTA and peer reviewed journals.
- To quantify the performance of these methods
- To consider alternative methods which may improve accuracy or reliability.
- To examine the utility of current methods when populations are heterogeneous.
- To investigate whether NICE processes may disincentivise the development of stratified therapies for heterogeneous populations.

3. Published works

This section outlines each paper and describes how they fit together to form a cohesive body of work.

3.1. Establishing current practice of survival analysis within health technology appraisals – Paper One

Aims

The primary aim of this paper [5] was to identify all recent cancer related technology assessments appraised by NICE and establish the methods used to model survival data and extrapolate, if necessary, across the patient population lifetime. NICE's evaluation of therapies on a cost-per-QALY basis is similar to that of PABC and CADTH, who are HTA bodies in Australia and Canada respectively. [48, 49] NICE's transparency in decision-making outcomes and its supporting methodology guides means it is very influential internationally. For example, it has established partnerships with Thailand's Health Intervention and Technology Assessment Program. [50] For these reasons, NICE technology appraisals were chosen as the focus for this paper.

The second aim of the paper was to establish how each appraisal considered and accounted for the uncertainty around estimates of life-years, e.g., through the exploration of different parametric models and parameter values.

Methods

A systematic search was conducted on the NHS Evidence Search Webpage, restricted to single technology appraisals of cancer therapies. The decision to focus on cancer appraisals was made to ensure that survival analyses would feature prominently, with a time-to-event outcome featuring as a key clinical outcome. The search was restricted to 2017, as this was the most recent full year at the time of performing the analysis and ensured that out-of-date methodology was not captured. For each appraisal, the publicly available NICE committee papers were used as the primary source of information which includes the written company submission and the EAG critique.

Results

A total of 28 appraisals were identified representing 22 distinct therapies across 16 types of cancer. Every appraisal used parametric models to extrapolate at least one time-to-event outcome for the economic model. In one submission, the company's preferred extrapolation method was to use splines, but in all other appraisals a standard parametric form was preferred (e.g. exponential or Weibull). In one appraisal the company modelled a time-to-event outcome using the population's survival time as estimated by a Kaplan-Meier curve, rather than fit a parametric model to the data. Where parametric models fitted poorly to data, a two-phase piecewise approach was taken, where parametric models were fitted to data beyond a certain point of follow-up, with the Kaplan-Meier estimator used to estimate survival prior to this point. The time horizons of the economic models ranged from 10 to 100 years, whilst the observed maximum follow-up periods from the key clinical trials ranged from 1.4 to 6.8 years. The mean for the percentage of the time horizon that had any observed follow-up was 12.4%. The reported median follow-up from each trial covered an average of 6.5% of the model time horizons.

All submissions reported using information criteria and the model plausibility when selecting an extrapolation. Most also considered visual fit to the data, and some compared extrapolations to external sources of data. All but three submissions considered alternative extrapolations to the preferred models, exploring some uncertainty in the choice of extrapolation. In nine appraisals the EAG agreed with the company's preferred choice of models for extrapolation.

Key messages and significance

This paper revealed the heavy reliance on extrapolation of parametric models for assessing the cost-effectiveness of a treatment and confirmed that information criterion played a big part in model selection, alongside model plausibility. It also demonstrated the large degree of subjectivity in the selection of a preferred survival extrapolation.

3.2. [Assessing practice of survival modelling in cost-effectiveness modelling outside of technology appraisals – Paper Two](#)

Aims

The primary aim of this paper [6] was to perform a systematic review of economic evaluations assessing licensed therapies for epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) negative advanced/metastatic non-small cell lung cancer (NSCLC). The project was funded by a small grant which dictated the disease area, however the paper was able to place additional focus on the modelling of time-to-event data compared to a typical systematic review of economic evaluations. Hence, the practice of obtaining survival estimates in published studies outside of NICE technology appraisals could be well documented, extending beyond the work in the first paper.

Methods

A literature search of key databases was performed according to the preferred reporting items for systematic review and meta-analysis (PRISMA) guidelines. The search was limited to articles published in the English language from 1 January 2001 until 26 Jul 2019, reflecting the time since NICE recommended docetaxel for NSCLC until the commencement of the review. All titles and abstracts were screened by two reviewers, and studies were included if they examined at least one relevant treatment recommended by NICE for NSCLC. Two reviewers each performed primary extraction from half of the studies and verified the other extractions. The

Consolidated Health Economic Reporting Standards (CHEERS)[51] and Philips[52] checklists were used for critical appraisal and quality assessment.

Results

Following the screening of 612 titles and abstracts, 54 publications were assessed at full text, with 30 eligible for inclusion in the systematic review, seven of which were NICE appraisals. Focusing here on the 23 publications that were not NICE technology appraisals, the reporting quality, as assessed by the Philips and CHEERS checklists, was generally high though these do not focus specifically on survival-related elements. Model structure was identical to that of models in most NICE cancer technology appraisals, featuring progression-free, post-progression and death health states, mostly modelled using either a Markov or partitioned survival model. These economic model types are described in more detail in Appendix C, alongside an overview of the survival related information for the studies. Some analyses used median survival times taken from observed follow-up or estimated the area under the Kaplan-Meier curve, and so did not fit any parametric models. The most common approach was to fit parametric models to estimate mean survival, either independently or assuming proportionality between comparators, with the model selection methods being generally consistent with technology appraisals. A small number of analyses used advanced techniques for extrapolation including cure models, piecewise models or restricted cubic splines. Two studies avoided any survival related parameters by assuming there was no difference between the technologies they were comparing. Factors associated with survival modelling such as hazard ratios, choice of extrapolation and cure proportions were consistently among the factors whose associated uncertainty had the largest impact on the cost-effectiveness outcomes. Many studies did not explore these areas of uncertainty comprehensively through scenario analyses, neither did they report any assumptions that were made as part of the economic analysis.

Key messages and significance

This paper established that the methods of extrapolating survival implemented in the literature were similar and at times far more basic than those used in NICE

technology appraisals. This further emphasised the need to generate evidence on the utility of these extrapolation methods for estimating life-years.

3.3. Assessing efficacy of parametric extrapolation – Paper Three

Aims

This paper [7] aimed to demonstrate the reliability of life-year estimates generated using parametric survival models. Firstly, the ability of parametric models to accurately estimate life-years from trial follow-up was examined. Secondly, the models preferred by each of AIC, BIC and log-likelihood were compared to establish whether one could be considered more appropriate for identifying optimal extrapolations from a selection of candidate models in technology assessments.

Methods

Monte Carlo simulations were used to model 12 scenarios which replicated follow-up of a single arm from four trials which provided key evidence in health technology appraisals reviewed by NICE. Single arm follow-up was chosen as it allows for a better understanding of the performance of the models to the data, and can be generalised to analyses of multiple arms by the basic properties of the mean and variance of a difference. Each trial was used to generate three different scenarios, where each assumed either an underlying exponential, Weibull or generalised gamma distribution of survival times. The trials had a range of sample sizes, follow-up length, event rates and clinical outcomes. Two trials had observed median survival. Eight parametric models (exponential, Weibull, log-normal, log-logistic, gamma, generalised gamma, Gompertz and generalised F) were fitted to complete and censored simulated data sets, and the corresponding life-years estimates were calculated, capped at the time horizon of each appraisal's economic model. AIC, BIC and log-likelihood values of the models were compared to select the optimal extrapolation and corresponding life year estimate according to each method. No assessment of plausibility of each model was made nor were hazard rates assessed as it is not practical to compare fitted and observed hazard rates for each

simulation. This represents occasions where judgement is made solely on goodness of fit statistics, which can be due to limited resources or a lack of information. It is also expected that the best-fitting models will have some correlation to those that are plausible and have hazard rates that are close to those observed.

Results

It was found that if the form of a fitted parametric model does not match the underlying distribution of the data, it is likely to produce biased estimates of life-years, even if follow-up is complete. Only data coming from a simple underlying distribution (exponential) could be reliably represented by the parametric models. BIC-preferred models had a lower mean-squared error than those preferred by AIC and log-likelihood when the median survival had been observed, however plausibility and underlying hazard shape assumptions should also be considered when selecting an extrapolation, rather than rely solely on any goodness-of-fit, to remove the possibility of an implausible extrapolation being chosen. When the underlying distribution contains multiple parameters, follow-up typical of a clinical trial at the point of appraisal did not contain enough information for the behaviour to be reliably captured when fitting a model. This was the case even if the model matching the underlying distribution was fitted. Models with the most parameters had the highest variation in their life-year estimates suggesting overfitting is a problem. Even when estimates of life-years were unbiased, variability was higher suggesting extrapolations cannot be considered reliable.

Key messages and significance

Applying AIC, BIC and log-likelihood without considering model plausibility can result in the selection of biased extrapolations as they only consider the fit to the observed period. This paper is the first step in evaluating the utility of present methodology for extrapolating survival data.

3.4. Considering the impact of assessing plausibility and exploring model averaging – Paper Four

Aims

This paper [8] aimed to extend the investigation of the third paper in two ways. Firstly, to consider the impact of including a plausibility assessment in the model selection algorithm on the performance of the parametric extrapolations and the methods of model selection. Secondly, to explore the utility of methods of model averaging at reducing the bias and improving the reliability of estimating life-years from a selection of parametric models.

Methods

Using the same simulation framework as the previous paper, the assessment of plausibility was introduced through a comparison to the true life-year estimate as calculated using the parameters from the underlying distribution. Models that produced estimates of life-years that had a difference greater than 25% from the underlying life-years were assumed as implausible and were removed from consideration, aiming to replicate the assessment of plausibility of extrapolations based on clinical expertise, though percentage differences of 15% and 50% were also considered. The possibility of the plausibility assessment being made on a biased prediction of the true life-years was also considered. A range of approaches to model averaging were included. Having identified the optimal model according to each of AIC, BIC and log-likelihood, different established thresholds of difference for AIC and BIC, that are usually used to indicate levels of distinction from the optimal model, were compared. Models that were within these thresholds of the optimal model for each simulation were averaged with equal weighting. Secondly, models were weighted according to their BIC, approximating Bayesian model averaging. The final averaging method considered was to take the mean average of all candidate models, with models weighted equally. The methods of model averaging were explored both with and without the plausibility assessment on the candidate models.

Results

Assessing the plausibility of candidate models improved accuracy and reliability of both the extrapolations and using on AIC, BIC or log-likelihood to select a preferred model. BIC remained the optimal method of single model selection in the scenarios considered. Model averaging was generally superior to relying on a single extrapolation, having a considerably lower mean-squared error in most scenarios and often outperforming even the true underlying distribution (Table A2, Appendix C). The main exception is when the underlying distribution is exponential where an exponential model was superior. On such occasions, relying on the single model preferred by BIC, or BIC based weighting, both performed well as the penalisation of the BIC favours models with few parameters, such as the exponential. Averaging offered little reduction in bias, as the estimates from the single model selection methods were already generally unbiased. Estimates from the generalised gamma model were usually unbiased, perhaps explained by its nested relationship with each of the true distributions, however it was associated with a large MSE as the limited data prevented precise estimation of multiple parameters. Model averaging was robust to a biased prediction of the true life-years by which candidate models were excluded for being implausible. The benefits of model averaging were more noticeable in the scenarios where median survival was not reached according to the Kaplan-Meier estimator.

Key messages and significance

This paper demonstrates the utility of current methods in extrapolating survival data, combining model fit with plausibility. It showed that there are benefits of averaging across multiple plausible candidate models compared to selecting a single model where there are uncertainties or complexities in the underlying hazard rate. Model averaging reduces the chance of obtaining an extreme value estimate of life-years, compared to selecting a single model, even once implausible models are removed from consideration. Selecting a single model may be appropriate when it is consistent with a hazard rate from mature data.

3.5. The problem of extrapolating from heterogeneous populations – Paper Five

Aims

This paper [9] extended the methodology of the previous two papers [7, 8] and applied it to the setting where the populations survival times did not come from a single parametric distribution, representing differing baseline prognoses or responses to treatment. It aimed to demonstrate the relationship between clinical estimates of effectiveness (hazard ratios) and efficacy estimates used in cost-effectiveness analyses (life-years) in heterogeneous populations. Additionally, this paper assessed the ability of parametric extrapolation when survival times came from a mixture of underlying distributions.

Methods

This simulation study modelled seven distinct scenarios where varying combinations of heterogeneity were represented in the intervention and control arms of a hypothetical randomised control trial. This paper used a similar approach to the previous simulation studies (papers 3 and 4). This time two arms were modelled for each scenario to explore the combined impact of varying kinds of heterogeneity occurring on either arm. An exponential distribution was used to sample the different survival times, whilst censoring times were estimated using a Gompertz distribution. Different combinations of exponential hazard rates modelled different treatment subgroup/complement treatment and prognostic effects. Each arm was divided into a subgroup and complement with their survival times generated independently. Current methodology was represented by fitting the eight candidate parametric models that were introduced in paper three, ruling out implausible models based on their predictions at 5 and 10 years. AIC and BIC were used to distinguish between the remaining plausible models. Other approaches that were considered included taking the average life-years from every plausible model and fitting separate models where a statistically significant treatment-subgroup interaction was detected within a trial arm.

Results

The paper's appendices contained a proof of how an exponential model fitted to censored follow-up from a heterogeneous population would always produce a biased estimate of life-years. The simulations showed that this bias was present when life-year estimates were obtained from the compared methods. Life-years for censored follow-up of heterogeneous populations tended to be underestimated. Due to the various combinations and presence of the heterogeneity across the scenarios, estimates of incremental benefit were either under- or overestimated.

The estimates of life-years from each sample of complete follow-up, without censoring, contained least but still considerable variability, suggesting that mature or complete follow-up of a trial cannot be relied upon to give an accurate prediction of life-years.

In a scenario where treatment benefit was overestimated, increasing the sample size slightly reduced the variability of estimates and bias, and increased the power to detect a treatment effect. Fitting separate models when significant interactions were detected without increasing sample size reduced bias but variability remained high. Increasing sample together with fitting separate models when significant interactions were detected resulted in relatively unbiased estimates for all methods and estimates with the least variation.

In the two scenarios where benefit was overestimated, model averaging was the best method when considering bias and mean-squared error simultaneously.

Key messages and significance

This study demonstrated the problems when current methods are used to extrapolate survival from trial data containing heterogeneous populations. These biased estimates of life-years are problematic as they result in unfair pricing of therapies according to NICE's existing method of assessing value, which then go on to become the reference point for future therapies.

3.6. Does NICE's appraisal process discourage the development of targeted therapies? – Paper Six

Aims

The aim of this paper [10] was to create a model which captured the key factors which influence the costs and effects of a treatment when its value is being assessed by NICE. This model was designed to explore the setting of a heterogeneous population where different subgroups have a different response to a treatment, and to investigate whether a pharmaceutical company and healthcare provider agree on their preference on whether an emerging therapy is developed as a stratified or unstratified treatment. It explores whether there may be another motivation to a company's decision not to investigate potential heterogeneous subgroups, in addition to the potential biased survival estimates identified in the previous paper. Whilst this paper does not explicitly model survival, it uses treatment benefit based on life-year estimates that were obtained from parametric extrapolation. The model developed in this paper was used to establish whether the preference of the healthcare provider/decision maker would match that of pharmaceutical company researching and producing the treatment.

Methods

A model was created capturing the major inputs that are considered when appraising a health technology from a healthcare provider's opinion. It was used to calculate a separate utility for the healthcare provider and for the pharmaceutical company and to deduce whether each stakeholder would rather a treatment with specific efficacy in a subgroup and complement population be developed as a stratified or unstratified treatment. The model included development costs of the treatment, alongside the conditional costs of developing a stratified therapy. For situations when the healthcare provider and pharmaceutical company might disagree, potential solutions were considered to align the preferences, which were (i) the healthcare provider raising its willingness-to-pay threshold for a stratified therapy, (ii) the healthcare provider paying an upfront lump sum contribution for the development of a stratified therapy, and (iii) the pharmaceutical company

incurring a penalty for the negative effects. The parameter values were obtained from a recent NICE appraisal of pembrolizumab for urothelial carcinoma.

Results

A region of misalignment of preferences of the healthcare provider and pharmaceutical company was identified using the values obtained from the appraisal. This occurred when the treatment had a positive effect in the subgroup but a negative effect in the complement and depended on the prevalence of the subgroup. The misalignment was driven by the costs associated with new treatment and the costs of developing and producing a biomarker test. The three solutions were successful at aligning the preferences, and the positioning and cost of the each depended on the subjective parameter of the true value of a year of health to the healthcare provider. The solutions were identical in terms of their alignment of the preferences, however differed on their influence on whether the pharmaceutical company is able to recover their development costs. The penalty term made this more likely, potentially discouraging drug development in other ways, whilst the other solutions made this less likely but came at an increased cost per life-year to the healthcare provider.

Key messages and significance

Under certain circumstances the current NICE framework may discourage pharmaceutical companies from developing stratified therapies, and companies may instead prefer to target a treatment to a wider population. Considering this alongside the potential for biased estimates of benefit when heterogeneity is present, assessors such as NICE should be cautious when appraising health technologies and look closely for potential differences in treatment effect that have not been identified by pharmaceutical companies, as permitted in the NICE methods guide. [46]

4. Discussion and recommendations

The works in this thesis have identified parametric models as a common method of extrapolating survival data in economic analyses and demonstrated their performance in a range of scenarios representative of NICE technology appraisals. They have demonstrated that BIC may be slightly superior to AIC when selecting a single model for extrapolation, and that care should be taken to avoid an extreme extrapolation when no information is known about model plausibility. Model averaging is well suited in such instances or when the best parametric form remains unclear; however, data immaturity can considerably reduce the performance of all considered extrapolation methods. The presence of heterogeneity can introduce bias into the parametric extrapolations, which may disincentivise the development of therapies for targeted subgroups alongside the additional costs associated with identifying these subgroups, if sufficient reward is not available.

Paper one was the first to give an overview of the methods used to extrapolate survival data in NICE technology appraisals since Latimer in 2013. [2] The size, variety and complexity of NICE appraisals mean it is difficult to standardise the approaches they use or to establish best practice. This paper achieved its aim to identify methods of extrapolation and accounting for survival-related uncertainty and generated valuable information as to the methods used alongside the characteristics of each study, including the supporting survival data and the extent of the reliance on extrapolations. It demonstrated that there was often disagreement between the different stakeholders over their preferred method of extrapolation, and that there were a range of approaches to exploring survival-related uncertainty. It serves as an important cornerstone for the remaining papers of this PhD.

A limitation of this paper was its short review period spanning 12 months, however this was selected on grounds of feasibility and in order to avoid identifying out-dated methods. This review could be improved by expanding the timescale covered and updating it to see what has changed since 2017. Changes in methods either from methodological advances or differing evidence bases may mean the results of

this review are outdated. It is possible that the challenges faced in extrapolating survival have evolved, as more appraisals rely on indirect comparisons and data from single-arm trials. A broader review would also allow investigation of the changes of methods over time, rather than the current cross-sectional approach.

The generalisability of this review could be extended if it included appraisals from other agencies that appraise technologies in similar ways and publicly release the documentation, such as the Canadian Agency for Drugs and Technologies in Health (CADTH).

The second paper provided a comprehensive overview of cost-effectiveness analyses for treatments of NSCLC, demonstrating the range of modelling techniques and assumptions used, including the methods of estimating survival benefit. The broader aim of this paper meant that this paper could not focus solely on methods of survival extrapolation, however it still found evidence that the methods used in literature overlap heavily with the methods that feature in NICE technology appraisals, allowing a wider generalisation of the future papers included in this thesis.

This paper included articles from as far back as 2002, meaning that some of the methods discovered may be considered outdated. The paper would have contributed more to this PhD if it had focused on the methods of extrapolating survival data, which could have allowed it to explore other disease areas as methods may vary in other populations and trials with different characteristics. Some of the studies in this review were also included in the review of NICE appraisals. Ideally this overlap would be avoided, perhaps through the production of a single review encompassing published literature and appraisals of health technologies.

The two review papers of this thesis identified the popularity of parametric models but since they were published there have been many papers that explore the potential of emerging methods including splines, mixture models, mixture cure models (MCMs), landmark models. A common motivation for these papers was that novel treatments such as immuno-oncological and gene therapies often are

associated with complex hazard rates that parametric models can struggle to capture. Hence, it is no surprise that in some of these papers parametric models are found to be inferior to the novel techniques, but there is no clear best approach and raises the question of whether parametric models can still be considered a part of current practice.

Some studies show that spline models are superior to parametric models when modelling data for immuno-oncological treatments [54, 55], whilst others show MCMs were superior to splines and parametric models. [56-59] Sometimes, the benefit of splines or MCMs was unclear, with no modelling approach clearly optimal. [60-62]

In contrast Roth et al. found that parametric models performed very similarly to MCMs [63], whilst Klijn et al. found a log-logistic parametric model was the closest fit to their extended follow-up. [33] MCMs have been found to produce biased estimates of survival when immature follow-up fails to accurately capture the cure proportion, and may not always be reliable even if their usage is clinically plausible. [64, 65]

A parametric model underpins modelling approaches such as mixture models, MCMs and piecewise models. Hence the findings of the simulations presented in this thesis may generalise to these more modern methods. Whilst these alternative approaches offer greater flexibility, they require either the estimation of more parameters compared to standard parametric models, or in the case of landmark models they use less follow-up, and so their benefits may not always be clear.

Almost all these studies used parametric models as the reference case, with a recent guide to selecting a flexible survival model for extrapolation recommending that parametric models are used as parsimonious reference case for more complex methods. [66] There is evidence that parametric models are still used in economic evaluations despite increasing awareness about alternative methods,[67, 68] even for immune-oncology and gene therapies. [69, 70] Parametric models are the only method of extrapolation explicitly mentioned in the updated NICE methods guide

(section 4.10.5), though it refers readers to TSD 21 when the proportional hazards assumption is violated when multiple sources of survival data require synthesising.

Whilst there may be times that parametric models are not appropriate, Kearns *et al.* recently described parametric models as “current practice” and it looks like they will remain popular for the foreseeable future. [71]

A review of NICE technology appraisals by Bell Gorrod *et al.* reported similar findings to the reviews of this thesis, despite covering a much longer period (2011-2017). [24] They found 91% of submissions used parametric models which were selected in consideration of their goodness of fit, with the EAG critical of the company’s preferred model in 71% of appraisals. Their recommendations were for a greater transparency and consistency in the application of survival methodology. Their review did not consider length of follow-up of the contributing data or the model time horizon, but this was considered in a review by Tai *et al.* which focussed the influence of immature data in NICE technology appraisals. [72] Tai *et al.* found that 41% of NICE appraisals between 2015 and 2017 used data described as immature by the EAG, and that this sometimes resulted in NICE approving the technology for a subgroup of the originally indicated population. They advocate for a review of past decisions when additional follow-up becomes available to ensure accurate estimation of survival benefit.

Whilst the majority of the literature focuses on methods used in NICE technology appraisals, a review by Grumberg *et al.* revealed that extrapolation using parametric models and piecewise modelling is very common when appraising immune-oncological technologies in France. [73] Furthermore, guidance from Haute Autorité de Santé recommends using a parametric model for extrapolation selected using information criteria meaning the results of this thesis have an international influence. [74]

Limitations of many of the papers that investigate the benefits of MCM or flexible parametric models are that they do so for the assessment of a single case study of a technology appraisal, or they use data that are not representative of those used in

a technology appraisal. Meanwhile, the simulation studies in this thesis considered a wide range of scenarios, all representative of technology appraisal.

Paper three was the first to demonstrate the utility of parametric models for extrapolation of survival data to estimate the lifetime benefit of a treatment. The paper is a helpful reference point to assess the utility of novel extrapolation methods and serves as a guide for decision-makers when establishing criteria for assessing whether a technology is cost-effective. The paper details the design of the simulations, demonstrating how to maximise the trial-based information to replicate follow-up which can be used as a template for future simulation studies of time-to-event outcomes.

A limitation of this paper is that it did not consider model plausibility and relied solely on AIC and BIC. The results of this paper are still informative in situations where no models can be ruled out on plausibility grounds. However, this paper would be improved if it had considered model plausibility. The results from the subsequent paper, which considered plausibility before selecting a single extrapolation model, were instead overshadowed by the methods of model averaging explored in that paper. The inclusion of flexible parametric models would have extended the scope and generated evidence on the comparable utility of these approaches.

NICE TSD 21 conducted a similar simulation study assessing the efficacy of parametric models alongside flexible parametric and cure models, without considering model selection or plausibility. [4] It considered settings with complex underlying hazard rates and showed parametric models generally performed poorly, however no model type performed well across all scenarios with all methods being capable of generating implausible extrapolations, particularly when heterogeneity was strongly present.

The main contribution of paper four is that it introduced and showed the potential benefit of averaging across multiple models for obtaining an estimate of long-term survival of a patient population. It compared these methods with the approach of

selecting a single model for extrapolation, reflecting how models are selected in NICE technology appraisals.

The averaging methods considered are novel and highly relevant to stakeholders involved in the appraisal of health technologies, who may not have access to any data and only have output of a set of candidate models fitted to the data. This is the situation faced by External Assessment Groups in NICE technology appraisals who often only have access to a company's partitioned survival economic model, and no patient level data.

Whilst this and the previous paper do consider 12 baseline scenarios, an improvement to the paper would be to consider additional scenarios where the underlying data might come from alternative distributions or where a treatment has a curative effect. The models considered in the paper were often nested relative to each other and to the underlying distribution, which may restrict the generalisability of the findings. Similarly, the underlying distribution was always included in the set of candidate models. The papers also only model the benefit of a single technology, and do not estimate a relative benefit, which is usually of interest to the decision maker. Neither paper includes background mortality, which is usually accounted for in NICE technology appraisals. Background mortality reduces the bias associated with models that are too optimistic and may improve the fit of the log-models in these simulations, however there are still plenty of diseases where background mortality is unlikely to influence the survival extrapolations as the modelled hazard rate always exceeds background mortality. The paper could have been more generalisable if it included alternative types of survival model such as piecewise, cure or flexible parametric models. The focus on parametric models is still informative as they underpin some of these emerging methods and are still commonly used in their original form.

The first two simulation papers of this thesis found a slight benefit of selecting models using BIC over AIC. Beca et al. also found BIC superior to AIC in their simulation study, however these results may be linked to the choice of underlying source distributions and may not extend to all situations. [75] Everest et al. only considered models preferred by AIC but reported that they produced biased

estimates of survival. [76] Overall, the performance of using information criterion to select a single extrapolation of trial follow-up does not seem reliable. However, it is not clear whether this is mostly attributed to the data or the methods of selection and extrapolation. Further work is needed to identify whether there is a point when data can be considered mature enough for extrapolation, through a combination of sample size, number of observed events and length of follow-up, that can be applied widely to technology appraisals.

The fifth paper demonstrated the links between different kinds of heterogeneity in treatment effects when assessed from the perspectives of a clinical- and cost-effectiveness assessor. This is an important but sometimes overlooked [53] consideration when assessing the cost-effectiveness of a health technology and is relevant whenever there may be a different response to treatment or baseline prognostic risk. The findings of this paper should motivate decision-makers to request more often further exploration of suspected heterogeneity and to interpret analyses of a heterogeneous population with caution.

The limitations of this paper are that it did not include a supporting case-study, which would have demonstrated the problem and utility of the solutions more clearly. The simulated scenarios considered were all based on survival times following an exponential distribution, whilst real survival time may have complex distributions which may influence the utility of the methods used. The paper could also have included alternative types of candidate model which may have coped better with a heterogeneous population, such as mixture or flexible models. It would be interesting to construct an economic model factoring in the different biases identified to explore the implications on the decision-making and technology pricing decisions.

The benefits of identifying subgroups where heterogeneity is suspected, and of fitting separate models accordingly has been shown, however sample sizes from trials may not be powered to detect such subgroups correctly. If subgroups cannot be identified, then methods such as mixture models, dynamic models, or flexible parametric models such as restricted cubic splines, may be better than standard parametric models at extrapolating for a heterogeneous population, however

research is needed to support this hypothesis. Mixture models and MCMs both assume there are two distinct patient subgroups present in a population. MCM assumes one of the subgroups is effectively cured, whilst mixture models just assume they come from two separate distributions. Their inclusion in the simulation of heterogeneity would have enhanced the paper, but these models have their limitations. Cislo *et al.* showed that mixture models can have convergence issues and require the specification of multiple initial values to ensure that the optimal value is obtained, [77] whilst the necessity of mature data for MCM has already been mentioned. Pharmaceutical companies may prefer to use cure models because of the known bias associated with MCM. Hence there are likely occasions where identification and separate modelling of subgroups will be a well-suited approach. Regular investigation of subgroups may lead to greater proactivity of pharmaceutical companies in their detection of optimal patient populations for their therapies.

The sixth paper shows how the current means of assessing the cost-effectiveness of a health-technology may discourage a pharmaceutical developer from identifying the specific subgroup of patients in whom the treatment is most effective. Importantly, this reveals that decision-makers such as NICE may need to consider incentivising the development of stratified therapies by paying more for them than a non-stratified therapy.

Paper six has limitations. The economic model relies on several unknown parameters, most notably the true value of a QALY to the healthcare provider. This parameter has considerable impacts of the degree of compromise achieved and the associated costs, yet remains a somewhat subjective and abstract value. The paper presents a simplified model of what is actually a very complex decision problem often spanning multiple populations and disease areas and approaches to appraising a health technology. This paper focused on NICE's appraisal process, and it is possible that the approaches of other decision makers may not have the same effect.

Antoñanzas *et al.* similarly assume some incentivisation may be required to encourage the development of stratified therapies and built a similar model which

implemented a policy where the healthcare provider only pays for the treatment of patients who respond well. [78] They say this can be effective motivation to pharmaceutical companies to identify key patient subgroups. Such an approach may be difficult to implement widely as some health outcomes can be subjective and it is unknown how patients would have performed on an alternative therapy. However, it may be preferable to the healthcare provider if they can avoid having to pay more. This could result in potential discrimination issues if pharmaceutical companies became too selective over which patients receive their therapies.

Although parametric models were identified as a current methodology in 2017, it is possible that their use is less common as alternative modelling approaches have become more popular. However, parametric models still act as a pivotal component in several more complex approaches, including mixture-cure and piece-wise models. Hence the findings of these papers are still relevant.

The simulation papers compared models to the true underlying distribution rather than the sampled dataset they were fitted to. This meant that the simulated datasets may not always have been representative of the truth, which may have made the models look better or worse. This is an interesting problem, as NICE appraisals generally assume the data from a trial are representative of the target population and for very mature data, the area under Kaplan-Meier plot is treated as the gold standard of life-year estimation. [5, 24] However, because of either differences in baseline characteristics or random chance, there may be occasions when the data are not representative of the target population. It would be interesting to repeat these simulation studies but assessing a model's performance by comparing it to the simulated data rather than the underlying truth, which may be more representative of current practice.

The investigation of heterogeneity was inspired by the experience of a wide variety of approaches observed in my experience of critiquing health technology submissions. Approaches now may be more standardised with the support of the newly published NICE methods guide.

Emerging techniques:

More recently, there is growing opportunity to factor in external sources of data when extrapolating survival outcomes. Historically external data would be used to validate a parametric extrapolation from a primary data source. [5, 24] Alternative approaches include those demonstrated by Wang et al. who investigated merging their dataset with the external data, but it was unclear whether the external data offered any improvement. [79] Pennington et al. explored a range of approaches including to assume a proportional hazard relationship between their internal and external data. [80] Aside from these methods, parametric models fitted to trial data generally only factor in external data through their use in a relative survival setting or through post-hoc adjustment such as treatment effect waning and background mortality.

Emerging techniques such as dynamic relative survival models and Bayesian multiparameter evidence synthesis enable incorporation of external data into the extrapolation. [81, 82] Blended survival models combine parametric models with Cox proportional hazards models to wane the treatment effect over time, which can also include external information. [83] NICE TSD 21 does not mention all these methods when considering approaches to using external data. The methods are also yet to feature routinely in NICE technology appraisals, possibly due to a combination of their complexity, recency and requirement for external data to be available in the correct format. These methods have the potential to provide improved extrapolations in cases where parametric models perform poorly.

From the works of this thesis, I draw the following recommendations for both the current appraisal of health technologies and for future research.

Recommendations for practice:

1. Take utmost caution when extrapolating without any consideration of plausibility and relying solely on information criterion, or when data are immature.
2. Consider averaging across all plausible models when optimal extrapolation is unclear.
3. Explore suspected heterogeneity and its potential impact on cost-effectiveness outcomes.
4. Consider HTA-related outcomes when designing clinical trials or other data-generating systems.
5. Re-evaluate treatment efficacy and value when trial follow-up is complete and/or when extended follow-up of real-world use is available.

Recommendations for future research:

6. Compare latest methods which replicates data and outcomes considered in NICE technology appraisals.
7. Explore alternatives to using lifetime horizons to reduce the dependence on survival extrapolations from inadequate data.

These recommendations are not all novel. Recommendation 3 is consistent with the recently published NICE methods guide which encourages clinical experts and other stakeholders to identify potential subgroups of interest. [46] The methods guide recommends identifying these in the scoping stage if possible, but permits later discovery and exploration.

Recommendation 4 is also made by Tai et al, [72] who state that even analyses with high levels of confidence in their cost-effectiveness should be revisited when more data is available.

Recommendation 6 means to combine the methods mentioned in TSD 21 [4] and the simulation style of this thesis where the steps of plausibility and model

selection are included to properly represent each approach, rather than just comparing classes of models and ignoring model performance within each class.

5. Conclusion

The works of this thesis have made significant contributions in the understanding and search for methods of obtaining an optimal survival extrapolation. Parametric models were identified and investigated using rigorous methodology to demonstrate their efficacy for estimating treatment benefit. Novel pragmatic methods of model averaging have been reported, and their benefits shown when data are too immature to reliably represent the true underlying behaviour. Complications in data such as heterogeneity and competing interests of stakeholders adds additional complexity to the already difficult task, however the works of this thesis informs current and emerging methods that seek to address these complexities.

Substantial challenges of predicting the future efficacy of a treatment remain but these papers have enhanced our understanding of the problem.

Acronyms

AIC: Akaike information criterion

ALK: Anaplastic lymphoma kinase

BIC: Bayes information criterion

CDF: Cancer Drugs Fund, UK

CHEERS: Consolidated Health Economic Reporting Standards

DSU: Decision Support Unit

EAG: External assessment group

EGFR: Epidermal growth factor receptor

EQ-5D: EuroQol 5 Domains

HTA: Health technology assessment

MCM: Mixture cure models

NHS: National Health Service, UK

NICE: National Institute for Health and Care Excellence

NSCLC: Non-small cell lung cancer

OS: Overall survival

PFS: Progression-free survival

PRISMA: Preferred reporting items for systematic reviews and meta-analyses

PSA: Probabilistic sensitivity analysis

QALY: Quality adjusted life year

RMST: Restricted mean survival time

TSD: Technical support document

References

1. Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, et al. Generalisability in economic evaluation studies in healthcare: a review and case studies. *Health Technol Assess*. 2004 2004/12//;8(49):iii-iv, 1-192.
2. Latimer NR. Survival Analysis for Economic Evaluations Alongside Clinical Trials—Extrapolation with Patient-Level Data: Inconsistencies, Limitations, and a Practical Guide. *Medical Decision Making*. 2013;33(6):743-54.
3. Latimer N. NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. Report by the Decision Support Unit. 2011.
4. Rutherford M, Lambert P, Sweeting M, Pennington R, Crowther MJ, Abrams K, et al. NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis. 2020.
5. Gallacher D, Auguste P, Connock M. How Do Pharmaceutical Companies Model Survival of Cancer Patients? A Review of NICE Single Technology Appraisals in 2017. *International Journal of Technology Assessment in Health Care*. 2019;35(2):160-7.
6. Gallacher D, Auguste P, Royle P, Mistry H, Armoiry X. A Systematic Review of Economic Evaluations Assessing the Cost-Effectiveness of Licensed Drugs Used for Previously Treated Epidermal Growth Factor Receptor (EGFR) and Anaplastic Lymphoma Kinase (ALK) Negative Advanced/Metastatic Non-Small Cell Lung Cancer. *Clin Drug Investig*. 2019 Dec;39(12):1153-74.
7. Gallacher D, Kimani P, Stallard N. Extrapolating Parametric Survival Models in Health Technology Assessment: A Simulation Study. *Medical Decision Making*. 2021;41(1):37-50.
8. Gallacher D, Kimani P, Stallard N. Extrapolating Parametric Survival Models in Health Technology Assessment Using Model Averaging: A Simulation Study. *Medical Decision Making*. 2021;41(4):476-84.
9. Gallacher D, Kimani P, Stallard N. Biased Survival Predictions When Appraising Health Technologies in Heterogeneous Populations. *PharmacoEconomics*. 2022 2022/01/01;40(1):109-20.
10. Gallacher D, Stallard N, Kimani P, Gökalp E, Branke J. Development of a model to demonstrate the impact of National Institute of Health and Care Excellence cost-effectiveness assessment on health utility for targeted medicines. *Health Economics*. 2022;31(2):417-30.
11. Raftery J, Powell J. Health Technology Assessment in the UK. *The Lancet*. 2013 2013/10/12//;382(9900):1278-85.
12. NICE. Guide to the processes of technology appraisal. 2018 [cited 27th January 2022; Available from: <https://www.nice.org.uk/process/pmg19/chapter/acknowledgements>
13. Drews J. Drug Discovery: A Historical Perspective. *Science*. 2000;287(5460):1960-4.
14. Zanders ED. Screening for Hits. *The Science and Business of Drug Discovery: Demystifying the Jargon*. Cham: Springer International Publishing; 2020. p. 187-97.
15. Zanders ED. Clinical Trials. *The Science and Business of Drug Discovery: Demystifying the Jargon*. Cham: Springer International Publishing; 2020. p. 241-65.
16. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*. 2020;323(9):844-53.
17. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*. 2016 2016/05/01//;47:20-33.
18. Dillon A, Landells LJ. NICE, the NHS, and Cancer Drugs. *JAMA*. 2018;319(8):767-8.
19. NICE. Guide to the methods of technology appraisal [PMG9]2013.

20. Group TE. EuroQol-a new facility for the measurement of health-related quality of life. *Health policy*. 1990;16(3):199-208.
21. Bernfort L, Gerdle B, Husberg M, Levin L-Å. People in states worse than dead according to the EQ-5D UK value set: would they rather be dead? *Quality of Life Research*. 2018 2018/07/01;27(7):1827-33.
22. Olsen JA, Lamu AN, Cairns J. In search of a common currency: A comparison of seven EQ-5D-5L value sets. *Health Economics*. 2018;27(1):39-49.
23. Garau M, Shah KK, Mason AR, Wang Q, Towse A, Drummond MF. Using QALYs in Cancer. *PharmacoEconomics*. 2011 2011/08/01;29(8):673-85.
24. Bell Gorrod H, Kearns B, Stevens J, Thokala P, Labeit A, Latimer N, et al. A review of survival analysis methods used in NICE technology appraisals of cancer treatments: consistency, limitations and areas for improvement. *Medical Decision Making*. 2019;39(8):899-909.
25. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ*. 1998;317(7172):1572-80.
26. ABPI. Stratified medicine in the NHS: An assessment of the current landscape and implementation challenges for non-cancer applications: ABPI; 2014.
27. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics*. 2008;26(9):733-44.
28. Charlton V. Does NICE apply the rule of rescue in its approach to highly specialised technologies? *Journal of Medical Ethics*. 2022;48(2):118-25.
29. Cookson R. Can the NICE "End-of-Life Premium" Be Given a Coherent Ethical Justification? *Journal of Health Politics, Policy and Law*. 2013;38(6):1129-48.
30. Mihram GA. *Simulation statistical foundations and methodology*: Academic press; 1972.
31. Wang Z, Bovik AC. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine*. 2009;26(1):98-117.
32. Latimer NR, Abrams KR. NICE DSU technical support document 16: adjusting survival time estimates in the presence of treatment switching. 2014.
33. Klijn SL, Fenwick E, Kroep S, Johannesen K, Malcolm B, Kurt M, et al. What Did Time Tell Us? A Comparison and Retrospective Validation of Different Survival Extrapolation Methods for Immuno-Oncologic Therapy in Advanced or Metastatic Renal Cell Carcinoma. *PharmacoEconomics*. 2021 2021/01/11.
34. Kearns BC, Stevenson M, Triantafyllopoulos K, Manca A. PCN444 DYNAMIC SURVIVAL MODELS FOR INCORPORATING EXTERNAL EVIDENCE WHEN EXTRAPOLATING OVERALL SURVIVAL: A CASE STUDY. *Value in Health*. 2019;22:S522.
35. Brookmeyer R, Crowley J. A Confidence Interval for the Median Survival Time. *Biometrics*. 1982;38(1):29-41.
36. Akaike H. Information theory and an extension of the maximum likelihood principle. *Selected papers of Hirotugu Akaike*: Springer; 1998. p. 199-213.
37. Volinsky CT, Raftery AE. Bayesian Information Criterion for Censored Survival Models. *Biometrics*. 2000;56(1):256-62.
38. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978;6(2):461-4.
39. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*. 2004;33(2):261-304.
40. Raftery AE. Bayesian model selection in social research. *Journal of Sociological Methodology*. 1995:111-63.
41. Bellmunt J, De Wit R, Vaughn DJ, Fradet Y, Lee J-L, Fong L, et al. Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *New England Journal of Medicine*. 2017;376(11):1015-26.

42. Alsaab HO, Sau S, Alzhrani R, Tatiparti K, Bhise K, Kashaw SK, et al. PD-1 and PD-L1 checkpoint signaling inhibition for cancer immunotherapy: mechanism, combinations, and clinical outcome. *Frontiers in pharmacology*. 2017;8:561.
43. Pembrolizumab Approved for First-Line Treatment in NSCLC. *Oncology Times*. 2016;38(23):29.
44. Paget M-A, Chuang-Stein C, Fletcher C, Reid C. Subgroup analyses of clinical effectiveness to support health technology assessments. *Pharmaceutical Statistics*. 2011;10(6):532-8.
45. Dias S, Sutton AJ, Welton NJ, Ades A. NICE DSU technical support document 3: heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011.
46. NICE health technology evaluations: the manual: National Institute for Health and Care Excellence (NICE); 2022.
47. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*. 2012;344:e1553.
48. Gandjour A. Willingness to pay for new medicines: a step towards narrowing the gap between NICE and IQWiG. *BMC Health Services Research*. 2020 2020/04/22;20(1):343.
49. Factors influencing HTA decisions differ between countries. *PharmacoEconomics & Outcomes News*. 2020 2020/11/01;866(1):15-.
50. Tantivess S, Chalkidou K, Tritasavit N, Teerawattananon Y. Health Technology Assessment capacity development in low- and middle-income countries: Experiences from the international units of HITAP and NICE. *F1000Res*. 2017;6:2119.
51. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. *International journal of technology assessment in health care*. 2013;29(2):117-22.
52. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health technology assessment (Winchester, England)*. 2004;8(36):iii-iv, ix.
53. Gallacher D, Armoiry X, Auguste P, Court R, Mantopoulos T, Patterson J, et al. Pembrolizumab for Previously Treated Advanced or Metastatic Urothelial Cancer: An Evidence Review Group Perspective of a NICE Single Technology Appraisal. *PharmacoEconomics*. 2019 2019/01/01;37(1):19-27.
54. 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 Medical Research Methodology*. 2020 2020/05/06;20(1):103.
55. Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, et al. Modelling the Survival Outcomes of Immuno-Oncology Drugs in Economic Evaluations: A Systematic Approach to Data Analysis and Extrapolation. *Pharmacoeconomics*. 2017 Dec;35(12):1257-70.
56. Bullement A, Latimer NR, Bell Gorrod H. Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study. *Value in Health*. 2019;22(3):276-83.
57. Federico Paly V, Kurt M, Zhang L, Butler MO, Michielin O, Amadi A, et al. Heterogeneity in Survival with Immune Checkpoint Inhibitors and Its Implications for Survival Extrapolations: A Case Study in Advanced Melanoma. *MDM Policy Pract*. 2022 Jan-Jun;7(1):23814683221089659.
58. Vadgama S, Mann J, Bashir Z, Spooner C, Collins GP, Bullement A. Predicting Survival for Chimeric Antigen Receptor T-Cell Therapy: A Validation of Survival Models Using Follow-Up Data From ZUMA-1. *Value Health*. 2022 Jun;25(6):1010-7.
59. Sussman M, Crivera C, Benner J, Adair N. Applying State-of-the-Art Survival Extrapolation Techniques to the Evaluation of CAR-T Therapies: Evidence from a Systematic Literature Review. *Advances in Therapy*. 2021 2021/08/01;38(8):4178-94.

60. Bansal A, Sullivan SD, Lin VW, Purdum AG, Navale L, Cheng P, et al. Estimating Long-Term Survival for Patients with Relapsed or Refractory Large B-Cell Lymphoma Treated with Chimeric Antigen Receptor Therapy: A Comparison of Standard and Mixture Cure Models. *Med Decis Making*. 2019 Apr;39(3):294-8.
61. Cooper M, Smith S, Williams T, Aguiar-Ibáñez R. How accurate are the longer-term projections of overall survival for cancer immunotherapy for standard versus more flexible parametric extrapolation methods? *J Med Econ*. 2022 Jan-Dec;25(1):260-73.
62. Felizzi F, Launonen A, Thuresson PO. Approximation of Long-Term Survival with Polatuzumab Vedotin Plus Bendamustine and Rituximab for Patients with Relapsed/Refractory Diffuse Large B-Cell Lymphoma: Results Based on The GO29365 Trial. *PharmacoEconomics - Open*. 2022 2022/07/28.
63. Roth JA, Yuan Y, Othus M, Danese M, Wagner S, Penrod JR, et al. A comparison of mixture cure fraction models to traditional parametric survival models in estimation of the cost-effectiveness of nivolumab for relapsed small cell lung cancer. *Journal of Medical Economics*. 2021 2021/01/01;24(1):79-86.
64. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. The Extrapolation Performance of Survival Models for Data With a Cure Fraction: A Simulation Study. *Value in Health*. 2021 2021/11/01/;24(11):1634-42.
65. Othus M, Bansal A, Erba H, Ramsey S. Bias in Mean Survival From Fitting Cure Models With Limited Follow-Up. *Value Health*. 2020 Aug;23(8):1034-9.
66. Palmer S, Borget I, Friede T, Husereau D, Karnon J, Kearns B, et al. A Guide to Selecting Flexible Survival Models to Inform Economic Evaluations of Cancer Immunotherapies. *Value Health*. 2022 Aug 13.
67. Li B, Alvir J, Stewart M. Extrapolation of Survival Benefits in Patients with Transthyretin Amyloid Cardiomyopathy Receiving Tafamidis: Analysis of the Tafamidis in Transthyretin Cardiomyopathy Clinical Trial. *Cardiology and Therapy*. 2020 2020/12/01;9(2):535-40.
68. Majer I, Kroep S, Maroun R, Williams C, Klijn S, Palmer S. Estimating and Extrapolating Survival Using a State-Transition Modeling Approach: A Practical Application in Multiple Myeloma. *Value in Health*. 2022 2022/04/01/;25(4):595-604.
69. Bullement A, Meng Y, Cooper M, Lee D, Harding TL, O'Regan C, et al. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available? *Journal of Medical Economics*. 2019 2019/03/04;22(3):205-14.
70. Hardy WAS, Hughes DA. Methods for Extrapolating Survival Analyses for the Economic Evaluation of Advanced Therapy Medicinal Products. *Hum Gene Ther*. 2022 Sep;33(17-18):845-56.
71. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. Comparing current and emerging practice models for the extrapolation of survival data: a simulation study and case-study. *BMC Medical Research Methodology*. 2021 2021/11/27;21(1):263.
72. Tai T-A, Latimer NR, Benedict Á, Kiss Z, Nikolaou A. Prevalence of Immature Survival Data for Anti-Cancer Drugs Presented to the National Institute for Health and Care Excellence and Impact on Decision Making. *Value in Health*. 2021 2021/04/01/;24(4):505-12.
73. Grumberg V, Roze S, Chevalier J, Borrill J, Gaudin AF, Branchoux S. A Review of Overall Survival Extrapolations of Immune-Checkpoint Inhibitors Used in Health Technology Assessments by the French Health Authorities. *Int J Technol Assess Health Care*. 2022 Mar 25;38(1):e28.
74. Unknown. Choices in methods for economic evaluation. *Public Health Assessment Haute Autorité de Santé*. 2020.

75. Beca JM, Chan KKW, Naimark DMJ, Pechlivanoglou P. Impact of limited sample size and follow-up on single event survival extrapolation for health technology assessment: a simulation study. *BMC Medical Research Methodology*. 2021 2021/12/18;21(1):282.
76. Everest L, Blommaert S, Chu RW, Chan KKW, Parmar A. Parametric Survival Extrapolation of Early Survival Data in Economic Analyses: A Comparison of Projected Versus Observed Updated Survival. *Value in Health*. 2022 2022/04/01;25(4):622-9.
77. Cislo PR, Emir B, Cabrera J, Li B, Alemayehu D. Finite Mixture Models, a Flexible Alternative to Standard Modeling Techniques for Extrapolated Mean Survival Times Needed for Cost-Effectiveness Analyses. *Value in Health*. 2021 2021/11/01;24(11):1643-50.
78. Antoñanzas F, Rodríguez-Ibeas R, Juárez-Castelló CA. Personalized Medicine and Pay for Performance: Should Pharmaceutical Firms be Fully Penalized when Treatment Fails? *Pharmacoeconomics*. 2018 2018/07/01;36(7):733-43.
79. Wang X, Adamson BJ, Briggs A, Tan K, Bargo D, Ghosh S, et al. Approaches for Enhanced Extrapolation of Long-Term Survival Outcomes Using Electronic Health Records of Patients With Cancer. *Value in Health*. 2022 2022/02/01;25(2):230-7.
80. Pennington M, Grieve R, der Meulen JV, Hawkins N. Value of External Data in the Extrapolation of Survival Data: A Study Using the NJR Data Set. *Value Health*. 2018 Jul;21(7):822-9.
81. Guyot P, Ades AE, Beasley M, Lueza B, Pignon J-P, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information. *Medical Decision Making*. 2017;37(4):353-66.
82. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. Dynamic and Flexible Survival Models for Extrapolation of Relative Survival: A Case Study and Simulation Study. *Medical Decision Making*. 2022;42(7):945-55.
83. Che Z, Green N, Baio G. Blended Survival Curves: A New Approach to Extrapolation for Time-to-Event Outcomes from Clinical Trials in Health Technology Assessment. *Medical Decision Making*. 2022;0(0):0272989X221134545.

Appendix A

Statements of authorship from co-authors

Paper 1:

Paper Title: How Do Pharmaceutical Companies Model Survival of Cancer Patients?
A Review of NICE Single Technology Appraisals in 2017

Paper Journal: International Journal of Technology Assessment in Health Care.

Paper Authors: Daniel Gallacher, Peter Auguste and Martin Connock

Published Date: 24 April 2019 (online)

Statement:

Daniel Gallacher generated the research idea and created the search strategy in discussion with Rachel Court. Daniel created the inclusion/exclusion criteria. Daniel and Peter separately reviewed the eligibility of papers identified in the search to systematically identify papers that adhered to the inclusion criteria. Daniel extracted and analysing the information from the eligible papers. Daniel drafted the manuscript and liaised with the publisher.

Co-Authors:

I hereby declare the above statement to be an accurate representation of Daniel Gallacher's contribution to the paper described at the start of the document.

<u>Name</u>	<u>Signed</u>	<u>Date</u>
Peter Auguste		07/03/2022
Martin Connock		07/03/2022

Paper 2:

Paper Title: A Systematic Review of Economic Evaluations Assessing the Cost-Effectiveness of Licensed Drugs Used for Previously Treated Epidermal Growth Factor Receptor (EGFR) and Anaplastic Lymphoma Kinase (ALK) Negative Advanced/Metastatic Non-Small Cell Lung Cancer

Paper Journal: Clinical Drug Investigation

Paper Authors: Daniel Gallacher, Peter Auguste, Pamela Royle, Hema Mistry and Xavier Armoiry

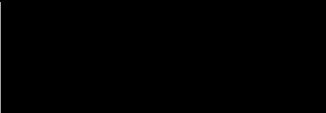
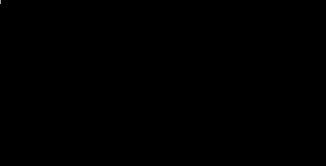
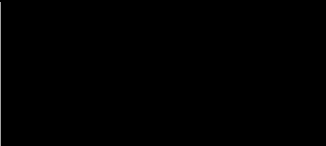
Published Date: 03 October 2019 (online)

Statement:

Daniel Gallacher was involved in the review, screening and data-extraction processes. Daniel also led the compilation and interpretation of results. Daniel led the write-up of the manuscript, producing the first draft, and liaised with journal on all matters concerning the paper.

Co-Authors:

I hereby declare the above statement to be an accurate representation of Daniel Gallacher's contribution to the paper described at the start of the document.

<u>Name</u>	<u>Signed</u>	<u>Date</u>
Peter Auguste		07/03/2022
Xavier Armoiry		07/03/2022
Hema Mistry		07/03/2022

Paper 3:

Paper Title: Extrapolating Parametric Survival Models in Health Technology Assessment: A Simulation Study

Paper Journal: Medical Decision Making

Paper Authors: Daniel Gallacher, Peter Kimani and Nigel Stallard

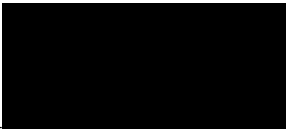
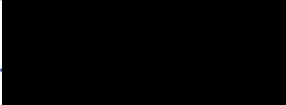
Published Date: 07 December 2020 (online)

Statement:

Daniel Gallacher generated the research idea and designed and implemented the first iteration of the simulation model. Daniel produced further iterations of the model and generated the simulation output. Daniel led the interpretation and description of the results and produced the first draft of the manuscript. Daniel was responsible for communicating with the journal over all aspects of the manuscript's publication.

Co-Authors:

I hereby declare the above statement to be an accurate representation of Daniel Gallacher's contribution to the paper described at the start of the document.

<u>Name</u>	<u>Signed</u>	<u>Date</u>
Peter Kimani		07-March-2022
Nigel Stallard		07-March-2022

Paper 4:

Statement of Contribution

Paper Title: Extrapolating Parametric Survival Models in Health Technology Assessment Using Model Averaging: A Simulation Study

Paper Journal: Medical Decision Making

Paper Authors: Daniel Gallacher, Peter Kimani and Nigel Stallard

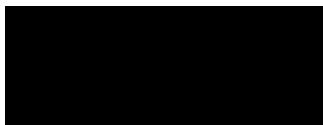
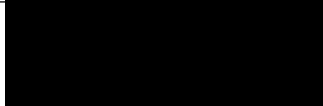
Published Date: 25 February 2021 (online)

Statement:

Daniel Gallacher generated the research idea and designed and implemented the first iteration of the simulation model. Daniel produced further iterations of the model and generated the simulation output. Daniel led the interpretation and description of the results and produced the first draft of the manuscript. Daniel was responsible for communicating with the journal over all aspects of the manuscript's publication.

Co-Authors:

I hereby declare the above statement to be an accurate representation of Daniel Gallacher's contribution to the paper described at the start of the document.

<u>Name</u>	<u>Signed</u>	<u>Date</u>
Peter Kimani		07-March-2022
Nigel Stallard		07-March-2022

Paper 5:

Paper Title: Biased Survival Predictions When Appraising Health Technologies in Heterogeneous Populations

Paper Journal: PharmacoEconomics

Paper Authors: Daniel Gallacher, Peter Kimani and Nigel Stallard

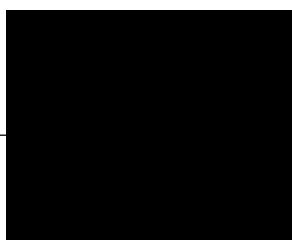
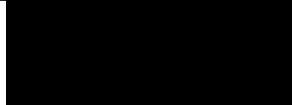
Published Date: 28 September 2021 (online)

Statement:

Daniel Gallacher generated the research idea and designed and implemented the first iteration of the simulation model. Daniel produced further iterations of the model and generated the simulation output. Daniel led the interpretation and description of the results and produced the first draft of the manuscript. Daniel was responsible for communicating with the journal over all aspects of the manuscript's publication.

Co-Authors:

I hereby declare the above statement to be an accurate representation of Daniel Gallacher's contribution to the paper described at the start of the document.

<u>Name</u>	<u>Signed</u>	<u>Date</u>
Peter Kimani		07-March-2022
Nigel Stallard		07-March-2022

Paper 6:

Paper Title: Development of a model to demonstrate the impact of National Institute of Health and Care Excellence cost-effectiveness assessment on health utility for targeted medicines

Paper Journal: Health Economics

Paper Authors: Daniel Gallacher, Nigel Stallard, Peter Kimani, Elvan Gökalp and Juergen Branke


Published Date: 25 November 2021 (online)

Statement:

Daniel Gallacher contributed to the research question and helped to design the decision model. Daniel created the model and identified suitable model inputs. Daniel conducted the primary analysis and all sensitivity and scenario analyses, generating the model outputs. Daniel produced the first draft of the manuscript and co-ordinated later versions. Daniel communicated with the journal on all matters regarding the manuscript's publication.

Co-Authors:

I hereby declare the above statement to be an accurate representation of Daniel Gallacher's contribution to the paper described at the start of the document.

<u>Name</u>	<u>Signed</u>	<u>Date</u>
Peter Kimani		07-March-2022
Nigel Stallard		07-March-2022
Juergen Branke		9 March 2022
Elvan Gokalp		10/03/2022

Appendix B

List of full works completed by the thesis author.

Title	Journal	Volume	Number	Pages	Year
Combination of Everolimus with Sorafenib for Solid Renal Tumors in Tsc2+/- Mice Is Superior to Everolimus Alone	Neoplasia	19	2	112-120	2017
Assessing the health economic agreement of different data sources	Stata Journal	18	1	223-233	2018
Pembrolizumab for previously treated advanced or metastatic urothelial cancer: an evidence review group perspective of a NICE single technology appraisal	Pharmacoeconomics	37	1	19-27	2019
Neurotrophins, cytokines, oxidative stress mediators and mood state in bipolar disorder: systematic review and meta-analyses	The British Journal of Psychiatry	213	3	514-525	2018
Ankle injury rehabilitation (AIR): a feasibility randomised controlled trial comparing functional bracing to plaster cast in the treatment of adult ankle fractures	Pilot and Feasibility Studies	5	1	1-8	2019

How do pharmaceutical companies model survival of cancer patients? A review of NICE single technology appraisals in 2017	International Journal of Technology Assessment in Health Care	35	2	160-167	2019
Derivation and internal validation of the screening to enhance prehospital identification of sepsis (SEPSIS) score in adults on arrival at the emergency department	Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine	27	1	1-13	2019
A Systematic Review of Economic Evaluations Assessing the Cost-Effectiveness of Licensed Drugs Used for Previously Treated Epidermal Growth Factor Receptor (EGFR) and Anaplastic Lymphoma Kinase (ALK) Negative Advanced/Metastatic Non-Small Cell Lung	Clinical Drug Investigation	39	12	1153-1174	2019
Rapid antigen detection and molecular tests for group A streptococcal infections for acute sore throat: systematic reviews and economic evaluation	Health Technology Assessment	24	31		2020
Systematic review and meta-analysis of the metabolic effects of modified-	Clinical Endocrinology	93	6	637-651	2020

release hydrocortisone versus standard glucocorticoid replacement therapy in adults with adrenal insufficiency					
Extrapolating Parametric Survival Models in Health Technology Assessment: A Simulation Study	Medical Decision Making	41	1	37-50	2020
Extrapolating Parametric Survival Models in Health Technology Assessment Using Model Averaging: A Simulation Study	Medical Decision Making	41	4	476-484	2021
Multivariate Generalized Linear Mixed-Effects Models for the Analysis of Clinical Trial-Based Cost-Effectiveness Data	Medical Decision Making	41	6	667-684	2021
Biased Survival Predictions When Appraising Health Technologies in Heterogeneous Populations	Pharmacoeconomics	40	1	109-120	2022
Development of a model to demonstrate the impact of National Institute of Health and Care Excellence cost-effectiveness assessment on health utility for targeted medicines	Health Economics	31	2	417-430	2022
Factors associated with attendance at screening for breast cancer: a systematic	BMJ Open	11	11	e046660	2021

review and meta-analysis					
Efficacy, safety, and dose-dependence of the analgesic effects of opioid therapy for people with osteoarthritis: systematic review and meta-analysis	Medical Journal of Australia	216		305-311	202 2

Appendix C

Contents:

Overview of survival analysis techniques used in this thesis.

- Kaplan-Meier survival function
- Exponential survival model
- Weibull survival model
- Log-normal survival model
- Log-logistic survival model
- Gompertz survival model
- Gamma survival model
- Generalised gamma survival model
- Generalised F survival model

Overview of economic model types related to the works in this thesis

- Markov model
- Partitioned survival model

Additional information from published works

Table A1: Survival related information extracted as part of paper 2.

Table A2: Comparison of methods of extrapolation across papers 3 and 4.

Kaplan-Meier survival function

The Kaplan-Meier estimator provides a way of accounting for censoring that occurs whilst following up the time-to-event outcome for a population, allowing follow-up of those who have not yet had the event to contribute information to the analysis rather than excluding them.

Let $S(t)$ be the survival function for event outcome death. In a population of n people, there are k unique event times each occurring at t_k . At the earliest event time, t_1 , d_1 of our population have the event. The probability of surviving beyond this time is $\frac{n_1-d_1}{n_1}$ or $1 - \frac{d_1}{n_1}$. This generalises to $1 - \frac{d_k}{n_k}$ for surviving each interval. But also accounting for previous intervals gives us: $S(t_k) = \left(1 - \frac{d_1}{n_1}\right) \left(1 - \frac{d_2}{n_2}\right) \dots \left(1 - \frac{d_k}{n_k}\right)$ which can be abbreviated as $\prod_{i=1}^k 1 - \frac{d_i}{n_i}$. This is the Kaplan-Meier survival function.

This can be represented graphically, where the survival function only decreases when an event occurs, as shown in Figure A1.

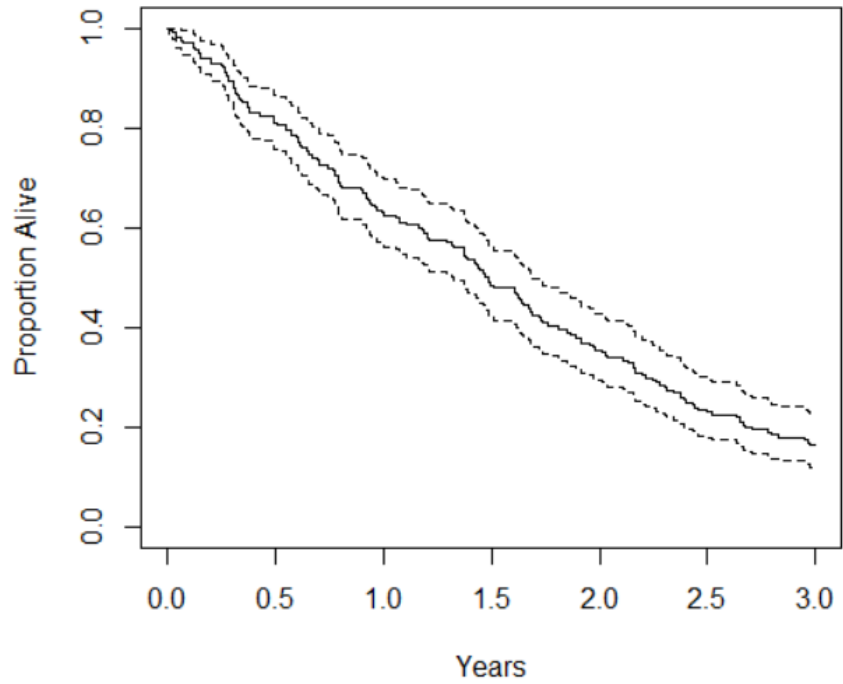


Figure A1: Example of a Kaplan Meier plot with 95% confidence interval.

Survival models

Parametric survival models are a group of models that can be used to represent the proportion of a population remaining event free over time, using a parametric form. They are useful because their defined parametric form means that their extrapolations each follow-up their own characteristic form, allowing simple predictions for the future survival of the patient population beyond the observed period. In this section, the survival functions of the parametric models used in the papers of this thesis are introduced briefly, alongside a visual representation of their survival and hazard forms.

Generally, the survival function can be written as:

The survival function where T is a non-negative random variable of event times is given by: $S(t) = P(T > t) = \int_t^{\infty} f(u) du$

where $f(t)$ is the probability density function given by: $f(t) = \lim_{\delta \rightarrow 0} \frac{P(t \leq T < t+\delta)}{\delta}$

The hazard function is the instantaneous failure rate at time t , and is defined as: $h(t) = \lim_{\delta \rightarrow 0} \frac{P(t \leq T < t+\delta | T \geq t)}{\delta} = \frac{f(t)}{S(t)} = \frac{-d}{dt} \log(S(t))$

Exponential survival model

$$\text{Survival function: } S(t) = \exp(-\lambda t)$$

where $\lambda > 0$

The survival and hazard profiles of a range of exponential distributions are shown in Figure A2, where the exponential model always has a constant hazard rate.

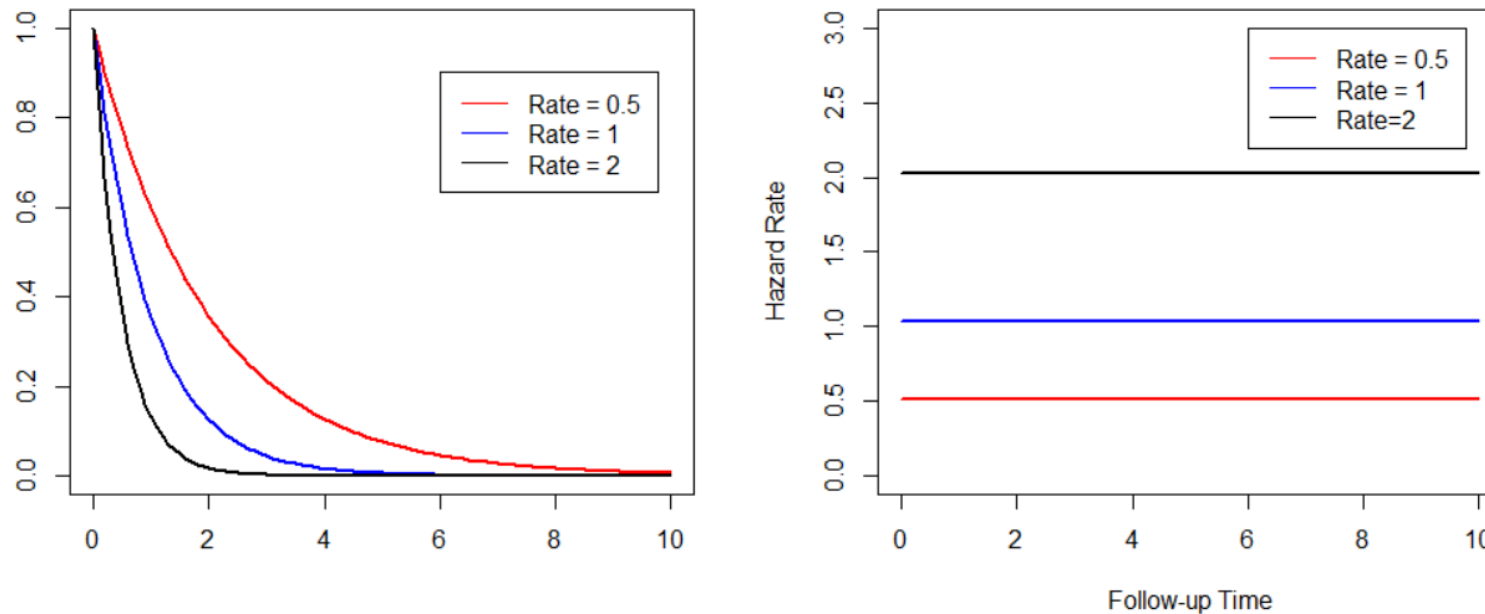


Figure A2: Survival and hazard plots for an exponential distribution

Weibull survival model

$$\text{Survival function: } S(t) = \exp\left(-\left(\frac{t}{\mu}\right)^a\right)$$

The survival and hazard profiles of a range of Weibull distributions are shown in Figure A3, where the Weibull model can have an increasing, decreasing or constant hazard rate.

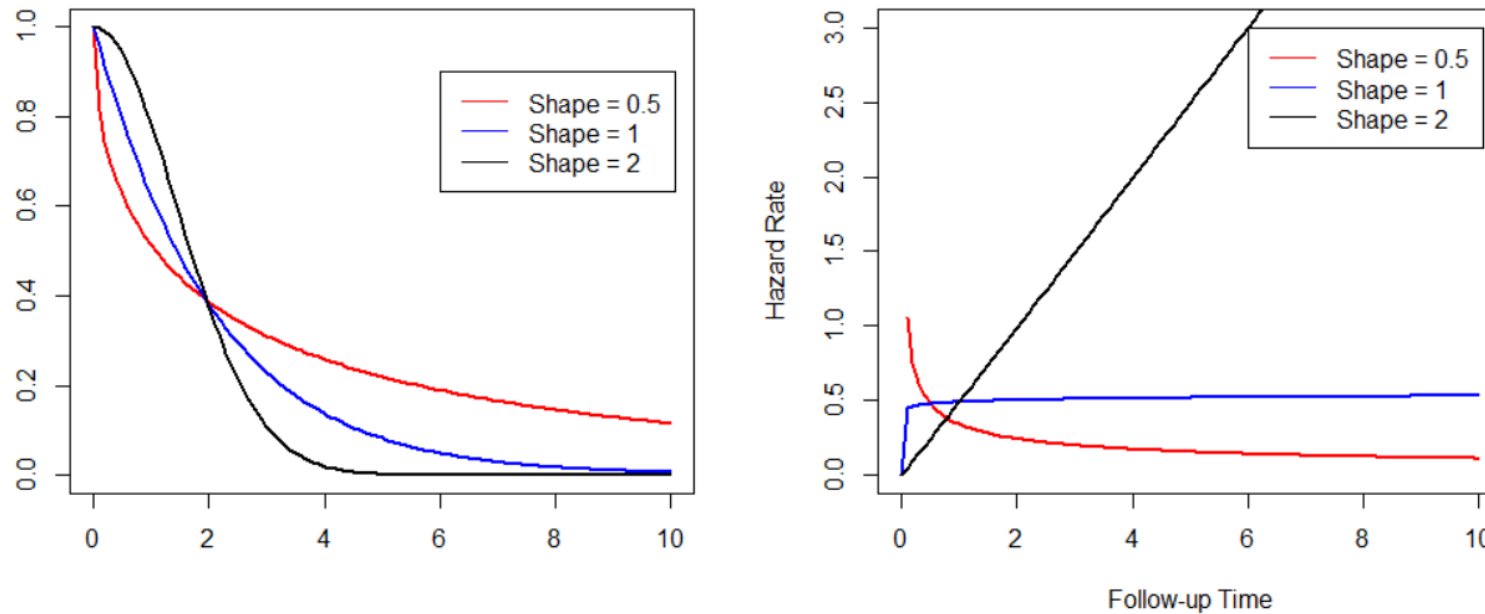


Figure A3: Survival and hazard plots for a Weibull distribution

Log-normal survival model

$$\text{Survival function: } S(t) = 1 - \int_0^t \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{\log(x-\mu)^2}{2\sigma^2}\right) dx$$

where $\sigma > 0$

The survival and hazard profiles of a range of log-normal distributions are shown in Figure A4 where the log-normal often has decreasing hazard rate long term with a high or low initial hazard rate.

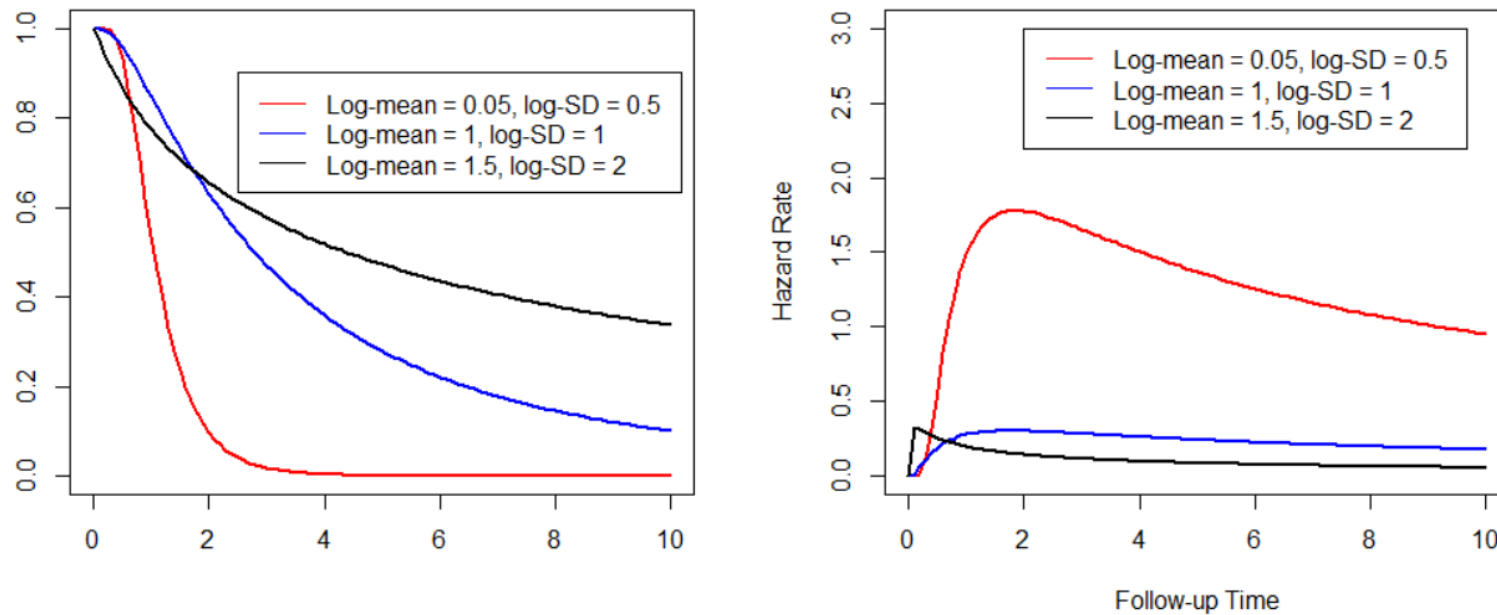


Figure A4: Survival and hazard plots for a log-normal distribution

Log-logistic survival model

$$\text{Survival function: } S(t) = \frac{1}{\left(1 + \left(\frac{t}{b}\right)^a\right)}$$

Where $a > 0$, $b > 0$

The survival and hazard profiles of a range of log-logistic distributions are shown in Figure A5 where the log-logistic often has decreasing hazard rate long term with a high or low initial hazard rate, like the log-normal.

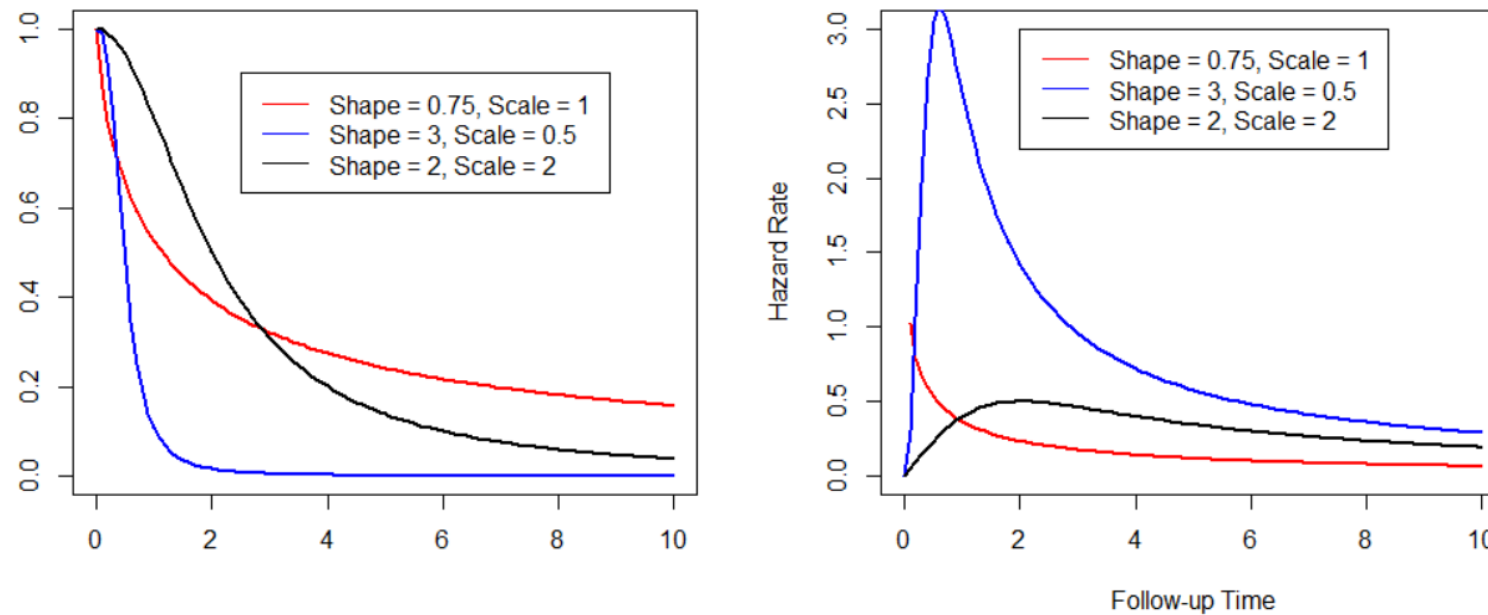


Figure A5: Survival and hazard plots for a log-logistic distribution

Gompertz survival model

$$\text{Survival function: } S(t) = \exp\left(-\left(\frac{b}{a}\right)(\exp(at) - 1)\right)$$

Where $b > 0$

The survival and hazard profiles of a range of Gompertz distributions are shown in Figure A6 where the hazard rate can either increase sharply or decrease to zero over time.

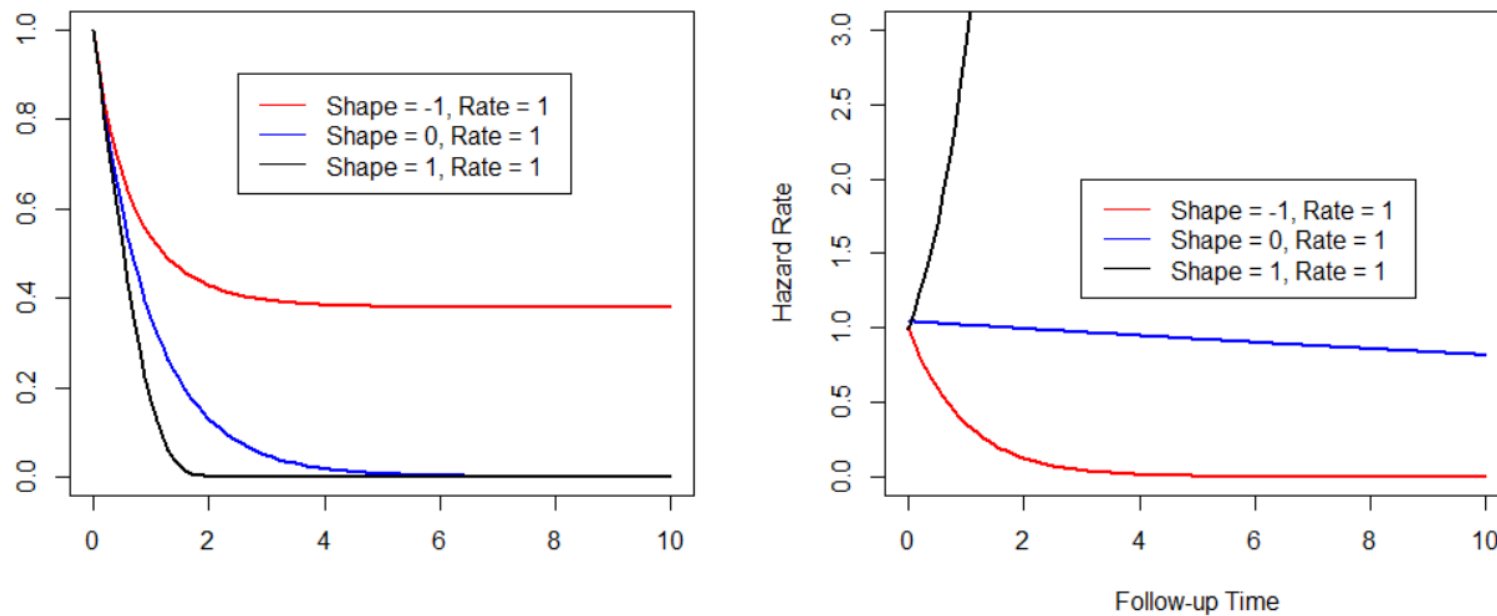


Figure A6: Survival and hazard plots for a Gompertz distribution

Gamma survival model

$$\text{Survival function: } S(t) = 1 - \int_0^t \frac{\exp(-\frac{x}{\mu})}{\mu^a \Gamma(a)} dx$$

where $\mu > 0$, $a > 0$ and $\Gamma(a)$ is the gamma function

The survival and hazard profiles of a range of gamma distributions are shown in Figure A7 where the hazard function can increase, decrease or remain constant over time.

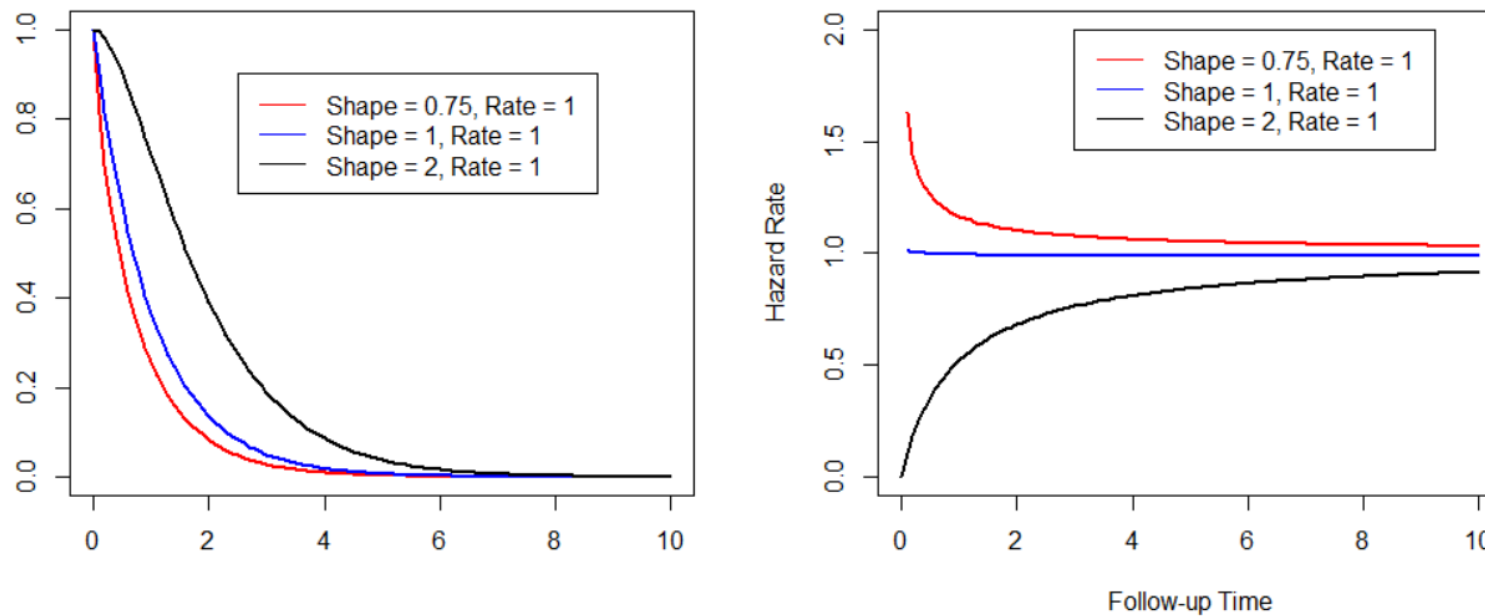


Figure A7: Survival and hazard plots for a gamma distribution

Generalised gamma survival model

$$\text{Survival function: } S(t) = \begin{cases} S_G\left(\frac{\exp(Qw)}{Q^2} \mid \frac{1}{Q^2}, 1\right) & \text{if } Q > 0 \\ 1 - S_G\left(\frac{\exp(Qw)}{Q^2} \mid \frac{1}{Q^2}, 1\right) & \text{if } Q < 0 \\ S_L(t \mid \mu, \sigma) & \text{if } Q = 0 \end{cases}$$

where $w = (\log(t) - \mu)/\sigma$, $\sigma > 0$; and $S_G\left(t \mid \frac{1}{Q^2}, 1\right)$ is the survival function of a gamma distribution with shape $a = 1/Q^2$ and scale = 1; and $S_L(t \mid \mu, \sigma)$ is the survival function of a log-normal distribution

The survival and hazard profiles of a range of generalised gamma distributions are shown in Figure A8 where the hazard rate can take a range of flexible forms.

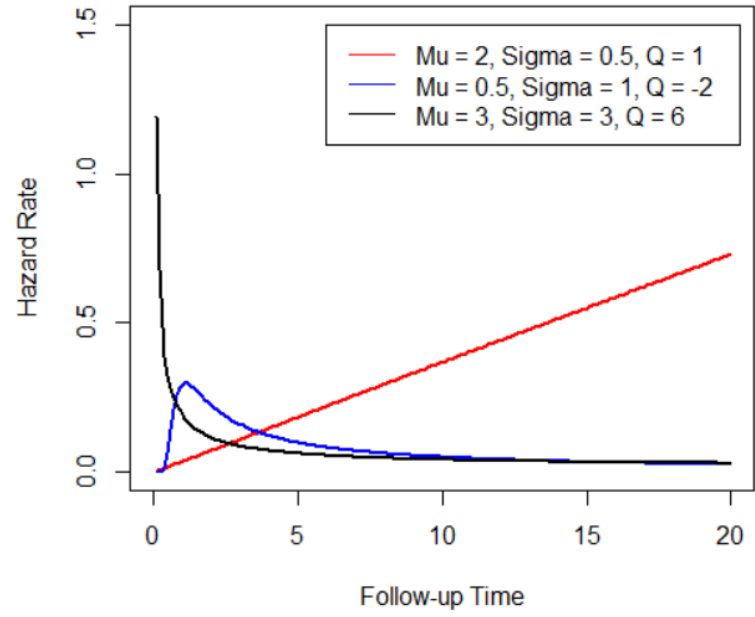
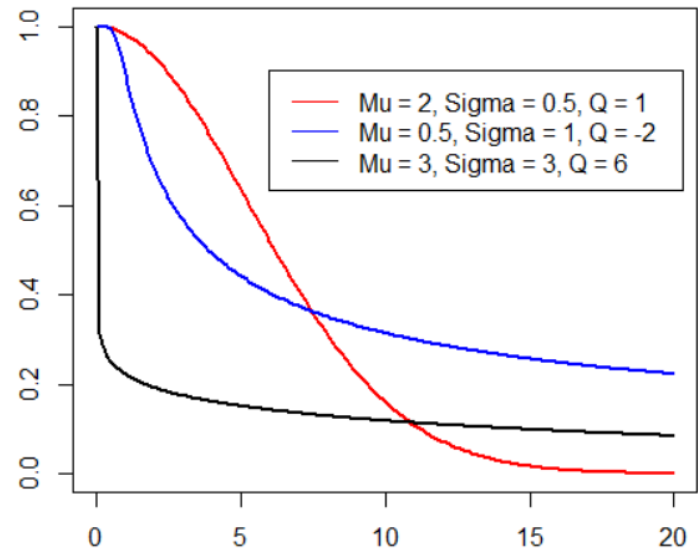


Figure A8: Survival and hazard plots for a generalised gamma distribution

Generalised F survival model

$$\text{Survival function: } S(t) = 1 - \int_0^t \frac{\delta \left(\frac{s_1}{s_2}\right)^{s_1} \exp(s_1 w)}{\sigma x \left(1 + \frac{s_1 \exp(w)}{s_2}\right)^{s_1 + s_2} B(s_1, s_2)} dx$$

Where $s_1 = 2(Q^2 + 2P + Q\delta)^{-1}$, $s_2 = 2(Q^2 + 2P - Q\delta)^{-1}$, $\delta = (Q^2 + 2P)^{\frac{1}{2}}$ and $w = \delta(\log(x) - m)/\sigma$ with $\sigma > 0$, $P > 0$, and B is the beta function.

The survival and hazard profiles of a range of generalised F distributions are shown in Figure A9 where the hazard rate can take a range of flexible forms.

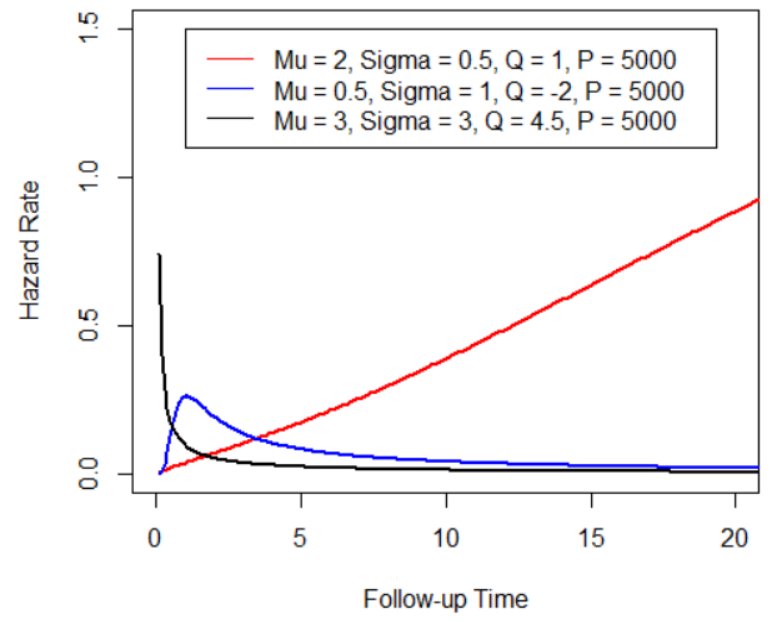
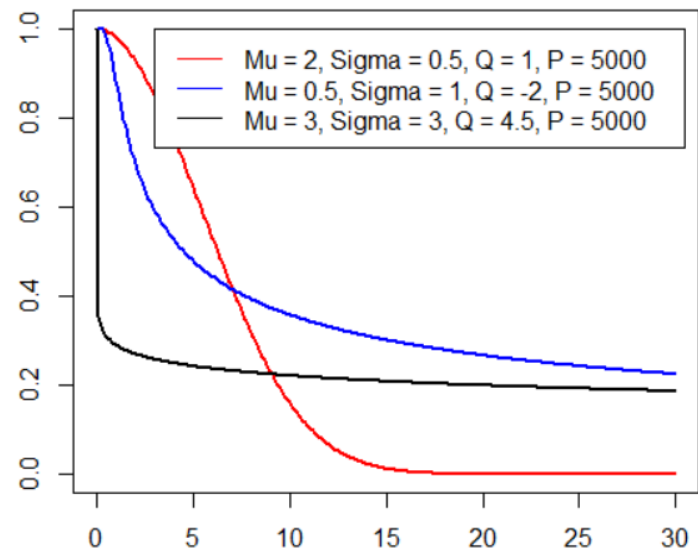


Figure A9: Survival and hazard plots for a generalised F distribution

Table A1: Summary of Survival Information Extracted for Paper 2: Systematic Review of Cost-Effectiveness Studies of Treatments for NSCLC

Paper/Source	Year	STA	How was survival modelled	How was model selected	Were other curves explored?	Were parameters included in PSA?
Araujo	2008	No	Parametric model - Weibull	Unclear	Yes	Unclear
Carlson	2008	No	No modelling, Equal PFS and OS was assumed	NA	NA	NA
Cromwell	2011	No	No extrapolation. Area under Kaplan Meier Curve	NA	NA	NA
Goeree	2016	No	Parametric models - proportional for OS	Goodness of fit statistics	No	No
Holmes	2004	No	No extrapolation, Area under Kaplan Meier Curve	NA	NA	NA
Leighl	2002	No	No extrapolation, Area under Kaplan Meier Curve	NA	NA	NA
Matter Walstra	2016	No	Constant hazard rate estimated from median survival	NA	NA	Yes
Pignata	2017	No	Parametric models - various	Best fitting models chosen	Yes	Unclear
Ramucirumab STA	2016	Yes	Parametric models, adjusted for covariates for OS (proportional), separately for PFS	Goodness of fit statistics, visual fit, plausibility	Yes	Unclear
Pemetrexed STA	2006	Yes	Constant hazard rate estimated from pooled data	NA	NA	NA
Nivolumab non squamous STA	2017	Yes	Parametric models were used, piecewise approach preferred by ERG	Goodness of fit statistics, plausibility	Yes	Yes

Nivolumab squamous STA	2017	Yes	Splines and parametric models were used	Goodness of fit statistics, visual fit, plausibility	Yes	Yes
Aguiar	2018	No	No extrapolation, Area under Kaplan Meier Curve	NA	No	No
Asukai	2010	No	Parametric model - exponential	Unclear	No	Yes
Cromwell	2012	No	No extrapolation. Area under Kaplan Meier Curve	NA	NA	NA
Greenhalgh	2015	No	Piecewise and spline models used, alongside Area under KM curve	Visual fit	Yes	Unclear
Huang	2017	No	Piecewise modelling was used	Goodness of fit statistics, visual fit	Yes	Unclear
Lewis	2010	No	No modelling, Equal PFS and OS was assumed	NA	NA	NA
McLeod	2009	No	Parametric model - exponential	Unclear	No	Unclear
Vergnenegre	2011	No	No extrapolation, Area under Kaplan Meier Curve	NA	NA	NA
Nintedanib STA	2015	Yes	Parametric models - various	Goodness of fit statistics, plausibility, external data	Yes	Unclear
Pembrolizumab STA	2017	Yes	Piecewise modelling was used	Goodness of fit statistics, visual fit, plausibility	Yes	Unclear
Atezolizumab STA	2018	Yes	Parametric models (piecewise models were considered)	Goodness of fit statistics, visual fit	Yes	Unclear
Bosch	2016	No	Used median survival	NA	NA	NA
Giurgis	2018	No	Used median survival	NA	NA	NA

Shafrin	2018	No	Parametric models proportional for OS, not PFS	Unclear	No	Unclear
Zhu	2018	No	Parametric models - various	Goodness of fit statistics	No	Unclear
Gao	2019	No	Parametric models - various	Goodness of fit statistics, visual fit	Yes	Unclear
Merino	2019	No	Used median survival	NA	NA	NA
Ondhia	2019	No	Parametric mixture cure models	Goodness of fit statistics	Yes	Unclear

Table A2: Comparing model averaging and model selection methods after plausibility to selecting the candidate models each time.

Drug/ Trial/ Outcome/ Source distribution	Measure	Results							
		Bayesian Model Averaging	Mean average of all models	Single Model Selection			Parametric Model (including implausible models)		
		Using BIC to estimate Bayes Factors		AIC	BIC	Log- likelihood	Exponential	Weibull	Generalised gamma
Dacomitinib/ ARCHER 1050/ OS/ Exponential	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	51.32 [+0.48, +1%] (44.45, 58.42) 51.24 [+0.40, +1%] 14.58 0.04 4.19 77%	52.25 [+1.41, +3%] (47.02, 58.54) 51.94 [+1.10, +2%] 14.58 0.04 3.55 83%	51.72 [+0.87, +2%] (42.10, 61.46) 51.63 [+0.78, +2%] 32.96 0.06 5.67 61%	51.04 [+0.20, +0%] (43.55, 59.27) 50.91 [+0.07, +0%] 22.40 0.05 4.73 71%	52.52 [+1.68, +3%] (39.69, 62.66) 53.78 [+2.94, +6%] 59.30 0.08 7.52 34%	51.03 [+0.18, +0%] (43.79, 59.09) 50.84 [-0.00, -0%] 21.93 0.05 4.68 72.5%	50.95 [+0.11, +0%] (41.79, 61.61) 50.46 [-0.38, -1%] 37.14 0.06 6.09 59.4%	51.33 [+0.48, +1%] (35.28, 69.81) 50.96 [+0.12, +0%] 115.16 0.11 10.72 31.2%
Dacomitinib/ ARCHER 1050/ OS/ Weibull	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	40.39 [+1.06, +3%] (34.71, 45.38) 40.51 [+1.18, +3%] 10.87 0.03 3.12 76%	40.16 [+0.83, +2%] (36.85, 44.18) 39.95 [+0.62, +2%] 5.50 0.02 2.19 90%	40.41 [+1.08, +3%] (32.58, 47.91) 40.60 [+1.26, +3%] 22.61 0.05 4.63 55%	40.45 [+1.11, +3%] (32.66, 47.95) 40.61 [+1.28, +3%] 22.60 0.05 4.62 55%	40.64 [+1.31, +3%] (32.17, 48.30) 41.02 [+1.69, +4%] 28.48 0.05 5.17 44%	52.27 [+12.93, +33%] (45.39, 59.73) 52.05 [+12.72, +32%] 186.18 0.04 4.35 1.2%	39.59 [+0.25, +1%] (34.40, 46.02) 39.19 [-0.14, -0%] 13.09 0.04 3.61 74.1%	40.90 [+1.57, +4%] (31.88, 54.52) 39.61 [+0.28, +1%] 52.26 0.07 7.06 45.3%
Dacomitinib/ ARCHER 1050/ OS/ Generalised gamma	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	37.67 [+1.94, +5%] (32.53, 42.17) 37.85 [+2.12, +6%] 12.18 0.03 2.90 66%	37.65 [+1.93, +5%] (34.32, 41.17) 37.54 [+1.82, +5%] 8.01 0.02 2.07 77%	37.32 [+1.60, +4%] (31.30, 43.56) 37.41 [+1.68, +5%] 17.57 0.04 3.87 56%	37.36 [+1.63, +5%] (31.37, 43.57) 37.44 [+1.71, +5%] 17.49 0.04 3.85 57%	36.98 [+1.26, +4%] (31.04, 43.81) 36.49 [+0.76, +2%] 18.15 0.04 4.07 57%	52.54 [+16.81, +47%] (45.69, 60.04) 52.36 [+16.64, +47%] 301.64 0.04 4.36 0.0%	40.00 [+4.28, +12%] (34.70, 46.72) 39.54 [+3.81, +11%] 32.08 0.04 3.71 46.7%	37.62 [+1.90, +5%] (30.77, 49.81) 36.14 [+0.41, +1%] 39.59 0.06 6.00 55.1%
Pembrolizumab / KEYNOTE 045/ PFS/ Exponential	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	5.62 [+0.02, +0%] (5.01, 6.29) 5.60 [+0.00, +0%] 0.15 0.00 0.39 85%	5.66 [+0.07, +1%] (5.07, 6.37) 5.64 [+0.04, +1%] 0.16 0.00 0.39 85%	5.63 [+0.03, +0%] (4.98, 6.40) 5.60 [-0.00, -0%] 0.18 0.00 0.42 82%	5.61 [+0.01, +0%] (5.00, 6.29) 5.59 [-0.01, -0%] 0.15 0.00 0.39 85%	5.80 [+0.20, +4%] (5.03, 6.73) 5.76 [+0.16, +3%] 0.30 0.01 0.51 70%	5.61 [+0.01, +0%] (5.01, 6.28) 5.60 [-0.00, -0%] 0.15 0.00 0.39 85.6%	5.62 [+0.02, +0%] (4.99, 6.33) 5.59 [-0.01, -0%] 0.17 0.00 0.41 83.6%	5.66 [+0.06, +1%] (4.97, 6.54) 5.61 [+0.01, +0%] 0.24 0.00 0.48 78.4%

Drug/ Trial/ Outcome/ Source distribution	Measure	Results							
		Bayesian Model Averaging	Mean average of all models	Single Model Selection			Parametric Model (including implausible models)		
		Using BIC to estimate Bayes Factors		AIC	BIC	Log- likelihood	Exponential	Weibull	Generalised gamma
Pembrolizumab / KEYNOTE 045/ PFS/ Weibull	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	5.76 [-0.08, -1%] (5.04, 6.59) 5.73 [-0.11, -2%] 0.23 0.00 0.47 77%	5.82 [-0.02, -0%] (5.13, 6.55) 5.81 [-0.03, -1%] 0.19 0.00 0.43 82%	5.85 [+0.01, +0%] (5.05, 6.81) 5.81 [-0.03, -1%] 0.28 0.01 0.53 73%	5.74 [-0.10, -2%] (4.98, 6.65) 5.70 [-0.14, -2%] 0.26 0.00 0.50 74%	6.05 [+0.21, +4%] (5.12, 7.07) 6.02 [+0.17, +3%] 0.39 0.01 0.59 64%	5.51 [-0.33, -6%] (4.87, 6.18) 5.49 [-0.35, -6%] 0.27 0.00 0.40 71.0%	5.87 [+0.02, +0%] (5.09, 6.75) 5.84 [-0.01, -0%] 0.26 0.01 0.51 76.1%	5.93 [+0.09, +1%] (5.03, 7.10) 5.85 [+0.00, +0%] 0.43 0.01 0.65 68.1%
Pembrolizumab / KEYNOTE 045/ PFS/ Generalised gamma	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	11.64 [-0.36, -3%] (9.22, 14.54) 11.52 [-0.48, -4%] 2.97 0.02 1.69 40%	11.83 [-0.16, -1%] (9.31, 14.54) 11.79 [-0.20, -2%] 2.65 0.02 1.62 45%	11.65 [-0.35, -3%] (9.20, 14.57) 11.54 [-0.46, -4%] 3.05 0.02 1.71 39%	11.61 [-0.39, -3%] (9.19, 14.54) 11.47 [-0.53, -4%] 3.05 0.02 1.70 39%	11.95 [-0.05, -0%] (9.24, 14.67) 11.96 [-0.04, -0%] 3.06 0.02 1.75 39%	5.16 [-6.84, -57%] (4.52, 5.84) 5.14 [-6.86, -57%] 46.95 0.00 0.40 0.0%	5.22 [-6.78, -56%] (4.53, 6.00) 5.19 [-6.80, -57%] 46.14 0.00 0.45 0.0%	12.58 [+0.58, +5%] (7.67, 19.23) 12.05 [+0.05, +0%] 13.38 0.04 3.61 26.3%
Pertuzumab/ APHINITY/ IDFS/ Exponential	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	379.33 [+2.49, +1%] (354.47, 404.69) 379.20 [+2.36, +1%] 249.41 0.16 15.60 98%	390.21 [+13.36, +4%] (358.88, 424.63) 389.42 [+12.58, +3%] 585.25 0.20 20.17 88%	385.92 [+9.08, +2%] (342.90, 458.14) 380.05 [+3.21, +1%] 1203.52 0.33 33.48 80%	377.32 [+0.48, +0%] (353.33, 401.88) 377.19 [+0.35, +0%] 241.63 0.16 15.54 98%	391.39 [+14.55, +4%] (298.19, 465.01) 401.66 [+24.82, +7%] 3186.65 0.55 54.55 33%	377.08 [+0.24, +0%] (353.71, 401.26) 377.13 [+0.29, +0%] 207.85 0.14 14.42 99.1%	373.33 [-3.51, -1%] (311.12, 429.13) 375.14 [-1.70, -0%] 1299.50 0.36 35.88 70.7%	377.52 [+0.68, +0%] (249.62, 489.12) 385.11 [+8.27, +2%] 7044.03 0.87 83.93 28.8%
Pertuzumab/ APHINITY/ IDFS/ Weibull	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	371.07 [+32.41, +10%] (328.01, 400.67) 373.89 [+35.24, +10%] 1520.06 0.22 21.67 46%	357.67 [+19.02, +6%] (325.18, 395.89) 355.85 [+17.19, +5%] 830.87 0.35 21.66 74%	361.77 [+23.12, +7%] (286.66, 403.79) 371.99 [+33.33, +10%] 1786.22 0.23 35.38 41%	378.24 [+39.58, +12%] (348.56, 406.75) 380.61 [+41.95, +12%] 2095.15 0.51 22.99 28%	349.91 [+11.25, +3%] (266.25, 417.13) 359.83 [+21.17, +6%] 2677.42 0.15 50.51 32%	381.15 [+42.49, +13%] (356.83, 405.38) 381.00 [+42.35, +13%] 2021.77 0.15 14.70 28.1%	336.85 [-1.81, -1%] (270.45, 400.14) 338.09 [-0.57, -0%] 1562.08 0.39 39.48 60.3%	340.62 [+1.97, +1%] (191.83, 476.34) 348.60 [+9.94, +3%] 9878.10 1.04 99.37 19.9%
Pertuzumab/ APHINITY/ IDFS/ Generalised gamma	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	370.36 [+36.83, +11%] (323.67, 400.54) 373.88 [+40.35, +12%] 1861.86 0.22 22.49 35%	355.84 [+22.31, +7%] (321.09, 392.76) 354.58 [+21.05, +6%] 974.82 0.36 21.84 69%	360.34 [+26.81, +8%] (285.67, 402.90) 371.40 [+37.87, +11%] 2008.97 0.24 35.92 34%	377.55 [+44.01, +13%] (346.54, 405.39) 380.34 [+46.81, +14%] 2506.15 0.50 23.85 17%	346.36 [+12.83, +4%] (262.95, 411.68) 357.66 [+24.13, +7%] 2672.00 0.15 50.08 30%	381.40 [+47.87, +14%] (357.92, 405.76) 381.06 [+47.53, +14%] 2504.79 0.15 14.59 39.82	336.71 [+3.18, +1%] (269.31, 400.80) 337.99 [+4.46, +1%] 1595.29 0.40 39.82	337.86 [+4.32, +1%] (185.86, 477.76) 343.81 [+10.28, +3%] 10242.31 1.05 101.12

Drug/ Trial/ Outcome/ Source distribution	Measure	Results							
		Bayesian Model Averaging	Mean average of all models	Single Model Selection			Parametric Model (including implausible models)		
		Using BIC to estimate Bayes Factors		AIC	BIC	Log- likelihood	Exponential	Weibull	Generalised gamma
							15.9%	58.9%	19.7%
Venetoclax/ MURANO/ OS/ Exponential	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	171.63 [+2.23, +1%] (148.60, 195.15) 171.72 [+2.31, +1%] 207.15 0.14 14.22 76%	175.04 [+5.63, +3%] (149.94, 199.66) 175.19 [+5.78, +3%] 270.34 0.15 15.45 65%	172.79 [+3.38, +2%] (143.26, 206.24) 172.08 [+2.67, +2%] 350.29 0.18 18.41 65%	170.12 [+0.71, +0%] (145.49, 194.56) 169.78 [+0.37, +0%] 222.36 0.15 14.90 75%	180.43 [+11.02, +7%] (133.61, 209.68) 188.18 [+18.77, +11%] 741.22 0.25 24.89 28%	170.04 [+0.63, +0%] (146.76, 193.83) 169.59 [+0.18, +0%] 202.54 0.14 14.22 77.5%	162.14 [-7.27, -4%] (107.63, 202.72) 166.07 [-3.34, -2%] 879.61 0.29 28.76 46.6%	174.89 [+5.48, +3%] (97.86, 216.69) 184.42 [+15.01, +9%] 1344.54 0.39 36.26 26.3%
Venetoclax/ MURANO/ OS/ Weibull	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	187.84 [-5.38, -3%] (162.23, 213.14) 187.76 [-5.46, -3%] 268.95 0.15 15.49 75%	192.64 [-0.58, -0%] (168.17, 212.42) 193.80 [+0.58, +0%] 178.98 0.13 13.37 85%	192.91 [-0.31, -0%] (156.98, 223.36) 194.39 [+1.17, +1%] 447.77 0.21 21.16 55%	181.90 [-11.32, -6%] (153.00, 220.18) 178.21 [-15.01, -8%] 546.31 0.20 20.45 48%	199.74 [+6.52, +3%] (159.19, 222.96) 203.96 [+10.73, +6%] 390.05 0.19 18.64 61%	170.46 [-22.76, -12%] (146.87, 193.85) 170.86 [-22.36, -12%] 722.87 0.14 14.32 40.3%	188.41 [-4.81, -2%] (153.92, 213.39) 191.17 [-2.05, -1%] 378.06 0.19 18.84 75.4%	191.52 [-1.70, -1%] (149.76, 220.05) 195.68 [+2.46, +1%] 583.31 0.28 24.09 49.0%
Venetoclax/ MURANO/ OS/ Generalised gamma	Mean (5%, 95%) Median MSE MCSE EmpSE % within 10%	187.64 [-2.63, -1%] (161.00, 213.27) 187.73 [-2.53, -1%] 258.79 0.16 15.87 75%	192.55 [+2.28, +1%] (166.72, 213.03) 193.85 [+3.59, +2%] 196.38 0.14 13.83 81%	192.32 [+2.06, +1%] (156.22, 223.56) 194.31 [+4.04, +2%] 461.87 0.21 21.39 54%	181.57 [-8.70, -5%] (152.44, 219.97) 177.84 [-12.43, -7%] 495.44 0.20 20.49 53%	199.65 [+9.39, +5%] (158.92, 223.46) 203.82 [+13.55, +7%] 449.67 0.19 19.02 53%	171.04 [-19.22, -10%] (147.51, 194.63) 171.21 [-19.05, -10%] 579.04 0.14 14.47 49.6%	188.41 [-1.85, -1%] (153.05, 213.66) 191.29 [+1.02, +1%] 377.25 0.19 19.34 73.4%	191.67 [+1.41, +1%] (148.58, 220.17) 195.91 [+5.64, +3%] 567.09 0.28 23.77 45.7%

Types of Economic Model

A simple common type of economic model is a Markov model. A visual representation of one is shown below in Figure A10.

This model consists of three health states: “Alive and Healthy”, “Alive and Unhealthy” and “Dead”.

The model will begin with a starting distribution of patients across the health states. Typically, everyone will start in the same health state, which in this case would be the “Alive and Healthy” box. Each arrow represents a transition probability that denotes the expected proportion of patients to move from one health state to another within a single cycle of the model. Patients can also remain where they are. A cycle length can be any period of time. Usually these transition probabilities are constant in a Markov model. Each health state is assigned an associated cost of care and a quality of life, which are then used to calculate total costs and QALYs for each treatment group and allow them to be compared.

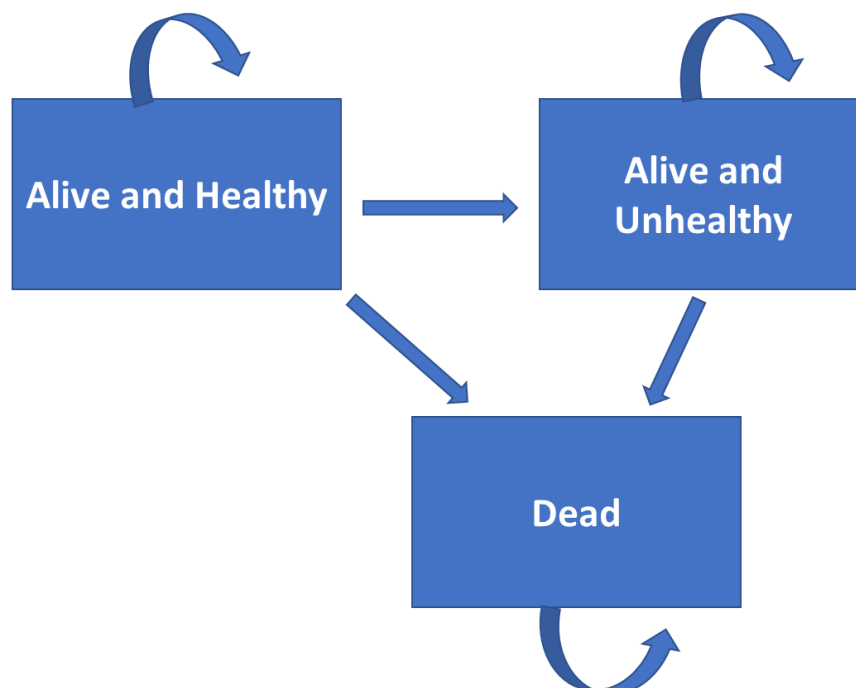


Figure A10: A visual representation of a Markov model

A partitioned survival model (PSM) is an alternative type of economic model, which is well suited to modelling diseases where the main health states can be reflected by time-to-event outcomes, such as PFS and OS. A PSM is shown below in Figure A11. The pink area under the red PFS curve represents the average time spent in the progression-free health state. The area above the blue OS line represents the average time spent in the death health state, whilst the light blue area in between the red and blue lines is the average time spent in the post-progression health state. The proportion in each health state can be estimated at any time and can be used to apply costs and quality of life values to estimate total costs and QALYs for each treatments allowing a comparison of their costs and benefits.

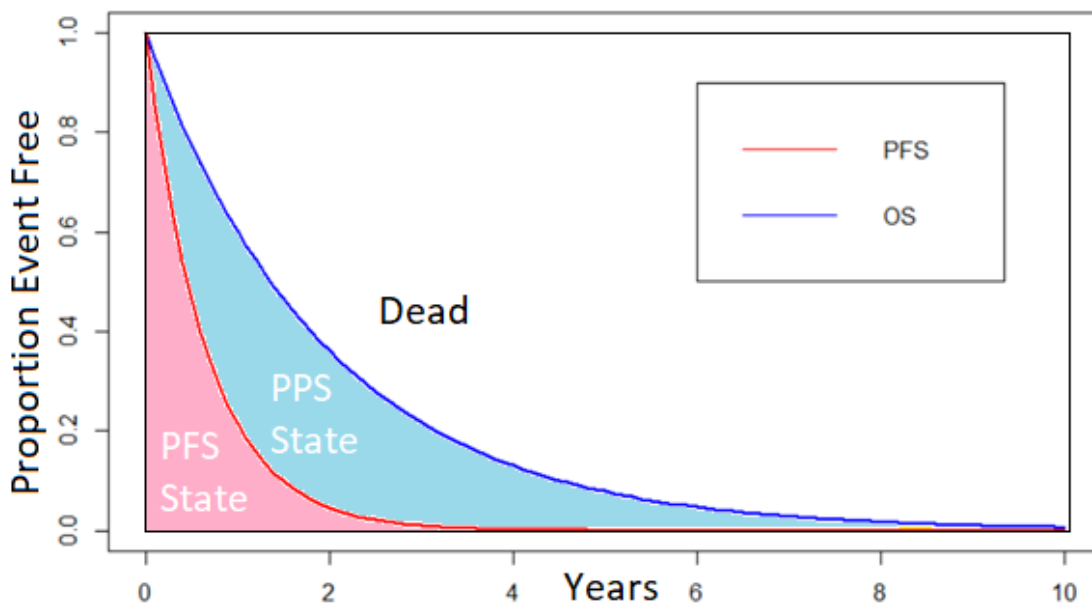


Figure A11: A visual representation of a partitioned survival model