Modelling treatment effect waning on survival outcomes to inform economic evaluation of healthcare interventions

MSc Statistics Dissertation

Student number: 22090905
University College London
September 2023
Word count: 11437

Supervised by
Prof Gianluca Baio (UCL)
Dr Richard Birnie (Lumanity)
Mr Matthew Hemstock (Lumantiy)

Abstract

Background. Survival extrapolation plays a crucial role in estimating the lifetime survival benefit of a treatment for health economic evaluation. Conventional survival extrapolation methods assume the long-term treatment effect to be consistent with the estimated treatment effect from the short-term trial. The advent of immune checkpoint inhibitors imposes a great challenge to this paradigm, as treatment effects of a wide range of immuno-oncology therapies wane slowly for several years after treatment discontinuation or disease progression. There has been a noticeable gap between the growing demand for modelling treatment effect waning and the underdeveloped methods available for addressing this issue. **Objective.** The author proposes the blending hazard method as a possible solution to modelling treatment effect waning in survival extrapolation. **Method.** The method fits separate survival regression models to short-term data from the internal RCT and long-term data from an external source and combine the two fitted hazard functions using a time-varying weight function for each treatment arm. The extrapolated hazard is initially dominated by the fitted internal hazard, then gradually approaches the fitted external hazard over a blending interval, and is finally dominated by the fitted external hazard. The weight function, characterised by a cumulative distribution function of a beta distribution, controls the rate of combining internal and external information during the blending interval. **Results.** The NICE TA366 on pembrolizumab for advanced melanoma not previously treated with ipilimumab is used as a case study to demonstrate practical implementation. Extrapolation and restricted mean survival time from the TA366 case study are shown to have a good consistence with the updated 7-year follow-up and perform a great improvement compared to the company's base case. **Conclusion.** The method maximises the utilisation of available information and possesses the adaptability to accommodate a wide variety of scenarios, thereby characterising associated uncertainty effectively.

Acknowledgements

I would like to express my deepest gratitude to my three supervisors: Prof Gianluca Baio has provided invaluable expertise and feedback from academic and NICE's perspectives that enhanced the robustness of my research; Dr Richard Birnie has been instrumental in offering crucial insights and consistently steering me in the right direction; Mr Matthew Hemstock has spent remarkable time and patience on addressing my questions, sharing relevant internal resources and contributing in-depth discussions throughout my work. This dissertation could not be completed without their extraordinary mentorship.

I would like to extend my thanks to Ms Emma Tyas (Lumanity) for her kind assistance in organizing my ideas and presenting thought-provoking insights at the early stage of my research. Thanks should also go to Ms Zhaojing Che (UCL) for her generous help in answering queries about her research and sending me useful reading materials.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

1.1.1 Health economic evaluation

Health economic evaluation, also known as health technology assessment (HTA), is the systematic evaluation of health interventions. The health interventions refer to a wide range of actions that aim to improve people's state of well-being, including physical and mental therapies, chemical-based and biological-based medications, diagnostic, surgical and rehabilitative procedures, etc. The systematic evaluation involves evidence-based cost-effectiveness analysis, which compares the cost with the benefits of interventions based on experimental evidence and real world evidence. Since the production of any health service is always subject to finite resources, especially for countries with some public funding of health care, governments need to intervene to optimise allocation and provision of health care via policies. Therefore, cost-effectiveness analysis is of great significance to inform policy-makers to prioritise interventions that provide the best value of money for the general population.

For most cost-effectiveness analyses, the effectiveness of an intervention is measured using utility that accounts for both people's length of life and perceived quality of life during that time. A commonly used utility measure is the quality-adjusted life year (QALY) which usually takes value between 0 and 1 (negative values of utility are possible but rare): 1 for expected one year of perfect health, 0 for death, other health states take values in-between. For each health state, the associated quality of life can be obtained from clinical measures or questionnaires, like the EQ-5D. The overall QALY for an intervention, is the sum of the number of years spent in some health states multiplied by their associated quality of life values. The use of utility provides a way to take different kinds of health effects to the same level, thus enables fair comparisons between interventions [1].

To compare the cost-effectiveness of two interventions (trt=0,1), we first calculate population average incremental benefit

$$\mathrm{E}[\Delta_e] = \bar{e}_1 - \bar{e}_0$$

and population average incremental cost

$$\mathrm{E}[\Delta_c] = \bar{c}_1 - \bar{c}_0$$

where $\bar{e}_{trt}$ and $\bar{c}_{trt}$ are population average benefit and cost for treatment trt, respectively. Then, intervention trt=1 can be represented by a point $(\mathrm{E}[\Delta_e], \mathrm{E}[\Delta_c])$ on the cost-effectiveness plane, as shown in Figure 1. The line passing the origin with slope k represents the willingness-to-pay (WTP) threshold, which is the maximum cost per incremental QALY that the healthcare provider (e.g. the government in the UK) wishes to pay. If the point lies below the line, treatment trt=1 is more cost-effective. If the point lies above the line, treatment trt=0 is more cost-effective. The most common case is that the point lies in the northeast quadrant of the cost-effectiveness plane, which means

treatment trt=1 is more effective but also more expensive. In this case, we can directly compute the Incremental Cost-Effectiveness Ratio (ICER)

$$\text{ICER} = \frac{\text{E}[\Delta_c]}{\text{E}[\Delta_e]}$$

which represents the incremental cost per QALY gained. This is to be compared with the willingness-to-pay. If ICER < WTP, treatment trt=1 is more cost-effective. If ICER > WTP, treatment trt=0 is more cost-effective.

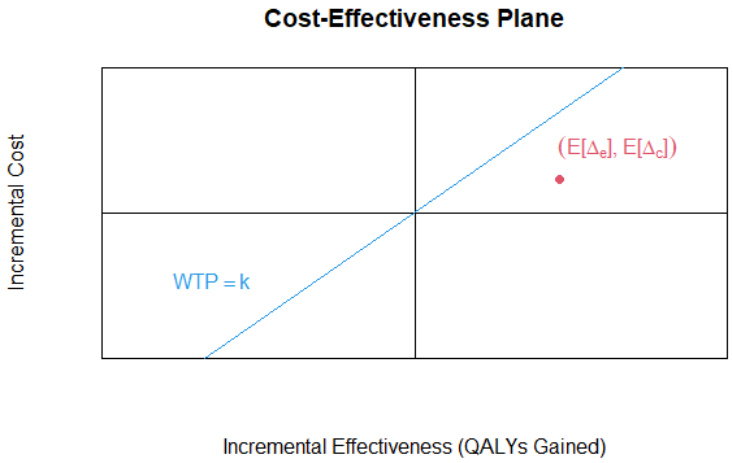

Figure 1 Cost-Effectiveness plane visualises the treatment comparison in cost-effectiveness analysis.

The uncertainty of the cost-effectiveness analysis can be modelled by the Cost-Effectiveness Acceptability Curve (CEAC), which quantifies the probability of a treatment being cost-effective with different values of willingness-to-pay (Figure 2). i.e. the probability that the ICER falls below the WTP threshold [2].

$$\text{CEAC} = \Pr(\text{treatment } trt\text{=1 is more cost-effective}|\text{data})$$

In England, the National Institute for Health and Care Excellence (NICE) is responsible for the cost-effectiveness evaluation of new medical interventions and provides evidence-based recommendations on whether interventions should be reimbursed by the National Health Service (NHS). In order to make new health technologies subscribed by the NHS, companies submit relevant evidence and analyses to NICE. NICE first commissions External Assessment Group (EAG, previously known as Evidence Review Group, ERG), an independent academic centre, to address strengths and weaknesses of the evidence. Then, NICE technical staffs organise company's submission and ERG report into committee papers for the appraisal committee. Based on evidence in the committee paper, appraisal committee members discuss publicly, make final decisions and publish Final Appraisal Determination (FAD) paper, which guides the use and funding of the assessed technology for the NHS.

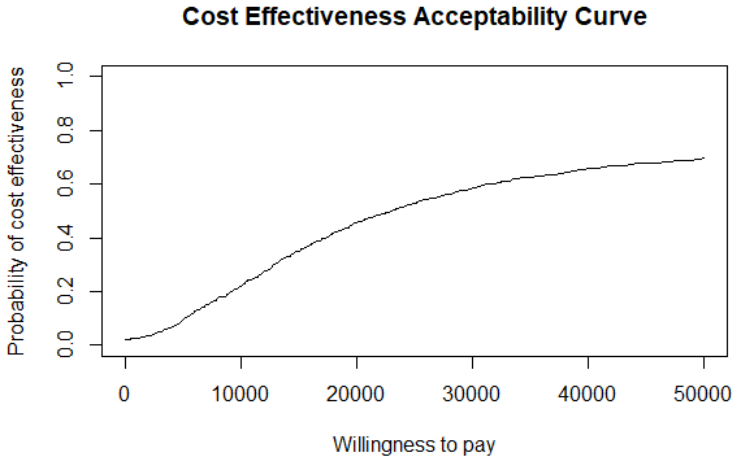

Figure 2 Cost-Effectiveness Acceptability Curve is the probability of a treatment being cost-effective, it quantifies the uncertainty for different willingness-to-pay in cost-effectiveness analysis.

1.1.2 Survival analysis and extrapolation

Survival analysis, also known as time-to-event analysis, models the time until the occurrence of some pre-defined event from a well defined time origin. Survival data are usually censored, which means that the event of interest does not occur during the observation period or the subject is lost to follow-up for some reasons. Survival data are usually recorded in the form of (t_i, d_i) for the $i = 1, \dots n$, where t_i is the time to event or censoring, d_i is the indicator for event, $d_i = 1$ for observed event, $d_i = 0$ for censoring. To visualise survival data, the Kaplan-Meier curve is often used to describe the observed survival probability over time. To produce meaningful analysis, parametric or semi-parametric models are fitted based on different assumptions. These models can be mathematically specified using a probability density function $f(t)$. This can be linked to the survivor function $S(t)$ that describes the estimated survival probability over time

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)du$$

and the hazard function $h(t)$ that describes the instantaneous risk of event over time [3]

$$h(t) = \frac{h(t)}{S(t)}$$

In health economic evaluation, since there are a wide range of interventions that aim to improve patients' survival, a great number of NICE technical appraisals involves survival analyses. This is typically reflected in oncology interventions, where studies focus on a number of survival endpoints. e.g. Overall survival (OS) - time from randomisation to death [4], progression-free survival (PFS) - time from randomisation until disease progression or death [4], distant metastasis-free survival (DMFS) - time from randomisation to the development of any distant metastasis or death [5], recurrence-free survival (RFS) - time from diagnosis or

randomisation to the first loco-regional or systemic recurrence [6], and duration of response(DoR) - time from randomisation to disease progression or death in patients who achieve complete or partial response [7], etc.

In NICE technical appraisals, as companies tend to obtain marketing authorisations and reimbursement approvals early, survival data are usually collected from interim analyses of randomised controlled trials as soon as the endpoint objectives are achieved. This means that the observed survival data are often immature due to short follow-up and can hardly cover the time horizon for health economic evaluations. A review of NICE single technology appraisals in 2017 indicates the mean proportion of time horizon supported by observed data is only 12.4 percent [8]. The immature data can be acceptable for clinical effectiveness analysis where hazard ratio, median survival time are of interest. However, it is a serious problem for cost-effectiveness analysis where we care about survival benefit not only in the trial follow-up time but also in the lifetime that beyond the observed duration [9] [10]. In other words, it is required to use short-term data to inform long-term survival in cost-effectiveness analysis. This task is generally referred to as the survival extrapolation. Since different models make distinct assumptions on the data generating process, the long-term survival estimates from various models can exhibit pronounced variation even if they are extrapolated from the same short-term data (Figure 3). Therefore, in order to inform sensible health economic evaluation, care must be taken to select appropriate models and make sure the extrapolation is realistic.

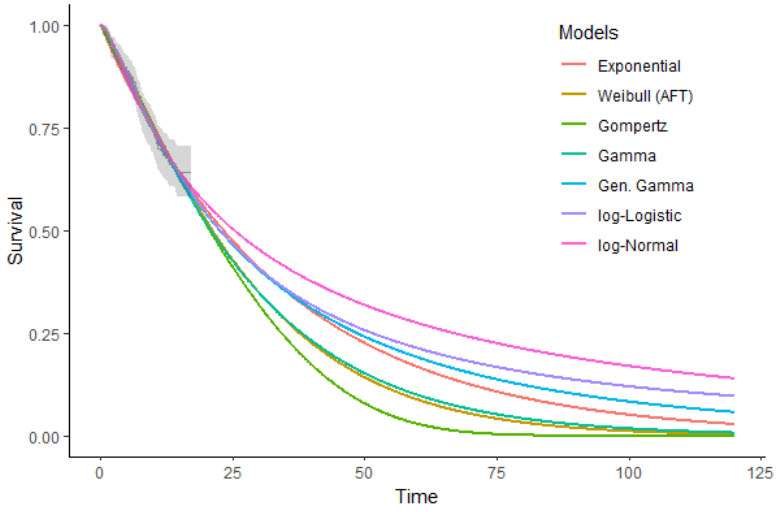

Figure 3 The long-term survival estimates from various models can exhibit pronounced variation even if they are extrapolated from the same short-term example data.

1.1.3 Overview of extrapolation methods

Historically, the most commonly used methods in NICE technical appraisals are the six standard parametric methods presented in the NICE Technical Support Document (TSD) 14: exponential, Weibull, Gompertz, loglogistic, lognormal, generalised gamma [8] [11]. The six standard parametric models assume different distributions of event over time and can generate different shapes of hazard

(constant, increasing, decreasing, increasing then decreasing, decreasing then increasing). It is generally recommended to consider all the six standard parametric models and compare them based on visual inspection, AIC, BIC, log cumulative hazard plots, clinical validity and external data. But, restricted by parametric forms, these models cannot capture complex hazard functions with more than one turning points [12].

Piecewise constant models are relatively more flexible as they segment the observed time into portions and assume hazard function is constant in each portion. This potentially allows better fit to the observed data than standard parametric models, however the extrapolation is restricted to have the same hazard as in the final observed portion and can be sensitive to the choice of segment points [13].

Landmark models divide observed time into two parts using a predefined landmark time point. The non-parametric Kaplan-Meier curve is used before the landmark and standard parametric models are fitted after the landmark. The two-piece construction is useful to reflect situation where treatment effect does not appear in the start of trial. The choice of landmark time point is often based on the time when treatment response is observed, but the exact time point can be controversial and constantly doubted by ERG.

Restricted cubic spline models [14] divide the observed time into portions using knots and fit a cubic polynomial to each portion in a way that the fitted hazard is continuous and smooth. The flexibility can be controlled by changing the number and positions of knots. Restricted cubic spline models are good at capturing complex shape of hazard function with many turning points [15], however, they are also at the risk of overfitting data. Since the end of Kaplan-Meier curve can present only a small number of patients, overfitting can influence extrapolation by capturing the trend appearing by chance at the end of the trial.

Fractional polynomial models fit observed data using polynomial functions where the exponents of polynomials can take non-integer values. This allows more flexibility to capture the trend of hazard than cubic or other standard polynomials. Whereas, careful consideration is required for selecting appropriate powers for fractional terms. Inappropriate choice of power may result in disastrously implausible extrapolation [16].

Mixture cure models assume observed survival is a mixture of cured survival and uncured survival [17]. The population consists of a cured group and an uncured group. For the cured group, it is assumed that the hazard will converge zero and the survival will forms a plateau. For the uncured group, conventional survival extrapolation methods listed above can be used. Mixture cure models are thought to be promising for modelling treatments with a cure effect. Nevertheless, the use of mixture cure method is only justified when the existence of a cure effect is supported by evidence and it is challenging to derive an accurate estimate of cure fraction from immature trial data [18].

All the extrapolation methods listed above can be categorised as trial-based extrapolation methods where only information from trial data is included for extrapolation. However, there are two common weaknesses for all trial-based methods.

One is that they extrapolate long-term hazard based on fitted short-term hazard, which implicitly assumes that the short-term survival trend will continue in the long term. This can be a naive assumption if we think about the potential

mechanism behind data of an immunotherapy for cancer: At the start of the trial, patients may not experience enough time to respond to the treatment but their disease progress, hazard increases rapidly. After a few months, a group of patients have responded to the treatment but patients who cannot respond or do not respond well continue to die, the increase of hazard slows down. Then, only good responders (long-term survivors) remain alive in the trial, hazard decreases. Over the next few years, the treatment effect wanes or the treatment is stopped, the decrease of hazard slows down. In the long term, hazard reaches a plateau or slightly increases due to other cause mortality and aging. It is extremely challenging to find a model that can capture such complex long-term changes in hazard by purely using short-term data. As for assessing model suitability, conventional measures like AIC, BIC and log cumulative hazard plots only evaluate goodness-of-fit to the observed data, none of them suggest the suitability of extrapolation. i.e. there is no external validation for the prediction.

Another common weakness of trial-based extrapolation is that they do not utilise all information about the treatment. Hard survival data from the short-term clinical trial are not the only source of information, external information from other relevant clinical trials and real world evidence can also help to inform extrapolation or reduce uncertainty. In theory, the optimal decision is only made when the extrapolation method incorporates all available information.

To address these problems in trial-based extrapolation, both NICE TSD 14 and NICE TSD 21 recommended to identify external data to help inform long-term extrapolation, where external data could come from other trials or registry data with longer follow-up in similar patient group [12] [13]. Over time, a large range of methods have been developed to incorporate external data into trial-based extrapolation [19].

One of the earliest method is to use external data as validation for selection of trial-based extrapolation models. Royston et al. [20] proposed a method to validate the prediction from fitted model on a secondary data set. TA135 [21] employed external data to assess the suitability of different alternative parametric models. Although this external data validation provides information for selecting appropriate trial-based models, it does not contribute to reducing the uncertainty in selected trial-based models. i.e. The external information is not incorporated into selected models.

A simple alternative is piecewise method (to be distinguished from previously discussed piecewise constant models), where the time horizon is cut into pieces and each source of data influences a piece of survival. Both TA357 [22] and Larkin et al. [23] presented 3-piece extrapolation: main trial Kaplan-Meier for the first 1-1.5 years; parametric curve fit to an external trial for the following several years; registry data and general population mortality for the final part. Nevertheless, this often implies sudden changes in fitted hazard at the places where pieces join together, thus the clinical plausibility is sometimes argued to be sceptical.

Another kind of method is to elicit external data by an informative prior distribution and combine it with observed data in a Bayesian way. Soikkeli et al. [24] first conducted parametric model selection on external data, then elicited an informative prior for the shape parameter in selected model based on the estimated shape parameter from external data. In this way, the distribution of the shape parameter can be updated by observed data from short-term trial.

This method performs well in reducing uncertainty in extrapolation and can easily update extrapolation when further follow-up data are available.

Excess hazard (relative survival) is a framework designed for incorporating general population mortality (GPM). It partitions all-cause mortality rate into disease-specific excess mortality rate and general population mortality rate. The general population mortality rate is a form of external data taken from population lifetables and population adjustment may be required in order to match the study population. NICE TSD 21 recommended to use excess hazard method to avoid extremely low hazard extrapolated from standard parametric models [13]. Sweeting et al. [25] demonstrated excess hazard method helps capture complex hazard and is relatively robust against lifetable misspecification.

Many other methods for incorporating external data are also potentially useful but relatively independent from each other. Che et al. [26] modelled observed data and external data separately and then blended them together using a weight function. Felizzi et al. [27] estimated cure fraction based on external data and applied the estimated cure fraction as an input of trial-based mixture cure model.

1.1.4 Treatment effect waning

With the development of immuno-oncology (IO), increasing number of immuno-oncology therapies have performed delayed treatment effects and long-term treatment effects that stay beyond trial follow-up [28] [29]. The biological mechanisms of immuno-therapies allow treatments to have deep and durable response and the treatment effects can wane slowly but still remain for several years after treatment discontinuation or disease progression. For example, a number of treatments have pre-specified stopping rules in administration, the treatment effects will gradually diminish for years after treatments are stopped, so it is not reasonable to assume their treatment effects at the end of trial will continue in the long term [30]. This long-term treatment effect imposes additional challenges for survival extrapolation in health economic evaluation, as it is unclear how the effect of a new treatment will evolve beyond the trial. Since trial data do not carry information about long-term effect, the only option is to resort to external data. Therefore, modelling treatment effect waning can be seen as a special case of incorporating external information into survival extrapolation.

The problem of treatment effect waning has raised people's concern in recent years. Although no NICE Technical Support Documents have provided guidance on how to model treatment effect waning (NICE TSD 21 only mentions treatment effect waning in the context of incorporating external data), some attempts have been made to take it into account. According to a review of treatment-effect waning for 7 immuno-oncology therapies in NICE technology appraisals [30], among 47 TAs before 2022, 12 TAs made waning assumptions in company submissions, ERG reports or committee comments. Another review focusing on nivolumab in NICE single technical appraisals [31] indicates 3 out of 10 TAs before November 2019 have incorporated treatment effect waning into committee accepted base case. Some published literature also expressed the concern to treatment effect waning. For example, Guyot et al. [32] stated that the hazard ratio of using cetuximab to enhance radiotherapy is expected to converge to one in 5 to 6 years, but there is a large uncertainty on this.

The most common method in NICE technical appraisals was to fit models based on proportional hazard assumption and set the hazard ratio to one at a specified time point that is assumed to have same hazard in both treatment arms. However, abrupt changes in hazard is an extremely strong assumption and can imply clinically implausible shape in survival curve. Only a small number of appraisals opted for proportional hazard methods that allow hazard ratio to gradually converge to one. Though gradual waning approaches seems more reasonable, the length and the rate of waning are unclear. Jackson [33] fitted a piecewise proportional hazard model to short-term data and assumed the log hazard ratio wanes linearly. However, this means the rate of waning completely depends on a strong and untestable assumption (log hazard ratio wanes linearly). A common limitation for these currently used methods are that they rely on proportional hazard assumption, which may fail to hold in practice.

Overall, there is a noticeable gap between the growing demand for modelling treatment effect waning and the underdeveloped methods available for addressing this issue. It is imperative to reevaluate appraisals when updated long-term follow-up are available and develop new data-driven methods that can characterise treatment effect waning in survival extrapolation for better decision making.

1.2 Presentation of the research

1.2.1 Blending hazard method

In this dissertation, the author considers the blended survival curves method from Che et al. [26] (a novel survival extrapolation method designed for incorporating external information) and amends that method to a slightly different method as a possible solution of modelling treatment effect waning in survival extrapolation. The original idea of blending comes from the prior work of Che et al., the author of this dissertation changes the method from blending survival functions to blending hazard functions. Since the new method is slightly different from the original one in both theory and results (not just a re-expression in hazard scale), to make statements clearly in the following sections, the new method is referred to as the blending hazard method, 'the author' is referred to as the author of this dissertation, though Che et al. still take full credit for the blending idea.

The idea of the blending hazard method is to first fit flexible survival models that provide reasonably good fit to short-term RCT data for each arm separately and fit another model that represents hazard in the long term for both arms from external data (e.g. from another trial about similar treatment for the same indication with longer follow-up or from registry database). Then, for each arm, the fitted hazard for internal RCT data (fitted internal hazard) and the fitted hazard for external data (fitted external hazard) are blended to a single hazard by a time-varying weight function, so that blended hazard is the same as fitted internal hazard in the initial period of time and gradually converging to fitted external hazard. The time-varying weight function is characterised by the cumulative distribution function of a beta distribution and can allow for different time and rate of blending, representing different time and rate of treatment effect waning.

The blending hazard method has three main advantages over other methods for treatment effect waning:

- All trial-based extrapolation methods can be incorporated into the blending hazard framework, there is no need to stick to proportional hazard assumptions;
- The method guarantees continuous hazard function throughout the waning period;
- Instead of completely relying on untestable assumptions (e.g. linear decrease in log hazard ratio), the rate of waning is partially data-driven (the rate is influenced by extrapolated shapes of fitted internal and external hazard).

1.2.2 Demonstrating case study

Selection of demonstrating case study: A good demonstrating case study for the method should satisfy :

- A NICE single technical appraisal (STA) that recommended the use of an immuno-oncology treatment as monotherapy, ideally with a waning assumption or a stopping rule, so that the treatment effect waning is perceived to exist;
- The Overall Survival (OS) Kaplan-Meier plot is available for the main trial of the technical appraisal to allow data digitisation. The focus is on OS because OS is the target of treatment waning (e.g. treatment are stopped after progression) and OS is often more immature and has more influence on health economic evaluation compared to Progression-Free Survival (PFS) [8];
- The OS Kaplan-Meier is very immature in the main trial, ideally the median survival is not reached for active treatment arm. This is because the problem of treatment effect waning particularly stands out in the scenario when data are very immature. When data are relatively mature, the blending hazard method still works, but its superiority to conventional methods may not be pronounced;
- The extrapolation is presented in the technical appraisal for ease of reevaluation and comparison.

The author reviewed all 28 NICE STAs involving pembrolizumab and nivoluamb for melanoma and non-small-cell lung cancer (NSCLC) before June 2023, as treatment effect waning is common in these appraisals. Among these TAs, 6 TAs were excluded for update and replacement by new TAs. The rest 22 TAs were screened based on inclusion of Recommended/Optimised categorisation and monotherapy. In the end, 10 STAs were identified for further inspection. By comparing these 10 STAs, TA366 [34] was considered to be the most appropriate case study, as its OS Kaplan-Meier plot of the main trial and extrapolation were presented as well as the OS data was immature (median survival was not reached for active treatment arm) in the main trial. Further details about the selecting process are shown in Figure 4 and Table 1.

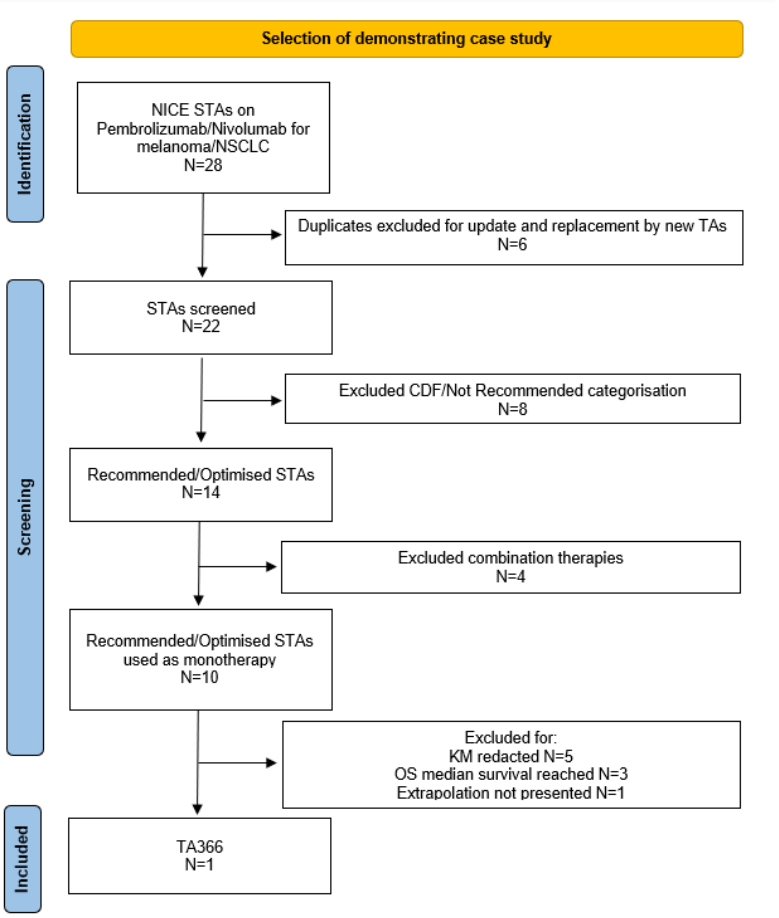

Figure 4 Flow diagram for selection of demonstrating case study.

Table 1: Comparison of screened case studies. TA366 was selected, as its OS Kaplan-Meier plot of the main trial and extrapolation were presented as well as the OS data was immature for active treatment arm in the main trial.

TA	Year	Categorisation	KM presented	Median survival not reached	Extrapolation presented
Pembrolizumab for melanoma					
TA357	2015/16	Recommended	Y	N	Y
TA366	**2015/16**	**Recommended**	**Y**	**Y**	**Y**
TA766	2021/22	Recommended	N	N	N
TA837	2022/23	Recommended	N	N	N
Pembrolizumab for NSCLC					
TA428	2016/17	Optimised	Y	N	Y
TA531	2018/19	Recommended	Y	N	N
Nivolumab for melanoma					
TA384	2015/16	Recommended	Y	Y	N
TA684	2020/21	Recommended	N	N	N
Nivolumab for NSCLC					
TA655	2020/21	Optimised	N	N	N
TA713	2021/22	Optimised	N	N	N

Selected demonstrating case study TA366: TA366 [34] recommended pembrolizumab, a humanised monoclonal antibody that targets the programmed cell death protein-1 immune-checkpoint receptor pathway, as an option for advanced melanoma not previously treated with ipilimumab. The main trial in TA366 is KEYNOTE-006, a phase III trial of pembrolizumab 10 mg/kg every 2 weeks or every 3 weeks, continued until disease progression or unacceptable toxicity, compared with ipilimumab 3 mg/kg every 3 weeks, continued for 4 doses. The trial was stopped after the second interim analysis as the primary endpoints (OS and PFS) had been met, where OS median follow-up was only 13.85 months and OS median survival were not reached for all arms.

The company's submission presented a 3-state Markov model comparing pembrolizumab (2 mg/kg every 3 weeks, maximum treatment duration 24 months) with ipilimumab, dabrafenib and vemurafenib in a 30-year time horizon. The OS extrapolation base case for pembrolizumab and ipilimumab were both estimated by KM data from KEYNOTE-006 in the first 13 months, followed by a model fit to the Schadendorf treatment-naïve population [35] (a pooled analysis of long-term survival data for people with melanoma treated with ipilimumab) from 13 months to 7 years, followed by a model fit to the Balch (2001) registry data [36]. The ERG criticised that the company's estimated hazard changed erratically at some time points of the time horizon and were not clinically plausible. The appraisal committee commented that there was not enough clinical evidence to believe pembrolizumab would provide a long-term survival benefit inconsistent with that shown in the ipilimumab trials. But the extrapolation for OS pembrolizumab were highly uncertain due to limitations in the use of external data.

The dissertation uses TA366 as a case study to demonstrate how the blended hazard method can be practically applied. With the availability of updated follow-up data, TA366 also offers an opportunity to reassess the company's approach and make a comparative analysis between the blended hazard method and the company's method.

2 Methods

2.1 Data collection

Data collection of the blending survival method involves: identifying the internal data source, identifying external data source, adjusting external data population for internal data and digitising pseudo individual patient data.

The internal data source refers to the main trial for clinical effectiveness analysis and cost-effectiveness analysis. The main trial provides the most accurate information on short-term survivals of treatments, but contains less information for the long term due to limited follow-up.

The external data source refers to other studies with longer follow-up that can inform the long-term survivals of treatments. Since the blending hazard method is designed for modelling treatment effect waning in health economic evaluations where treatments are assumed to eventually have the same hazard, the external data source needs to be a trial that informs long-term survivals of all treatments. When a novel treatment is compared with an old treatment, the external data source can be a trial on the old treatment with long follow-up, as old treatments are usually understood better with some long-term trial data available. When no trial data on the old technology are available, the external data source can be other trials on the same treatment for the same disease but different treatment line or on similar treatment for the same disease. Another option is to use registry data as external data, like SEER or AJCC staging database. Registry data usually show survival performance in a much longer term than trial data and inform mortality for general population or a specific group, but can be less relevant to the short-term trial population.

It is common that there is a divergence in population between the internal data and the identified external data in terms of age, gender, ethnicity, disease stages, prior therapy, etc. In order to provide the most accurate information on long-term survival, population adjustment methods might be needed to ensure the external data population matches with the internal data (e.g. propensity score, g-computation).

Finally, since individual patient survival data is required by further analysis but the reported Kaplan-Meier does not directly provide that data, the reported Kaplan-Meier plots should be digitised by the computer and then the Guyot algorithm [37] can be applied to map those digitised Kaplan-Meier plots back to the pseudo individual patient data. The data can be recorded in the form of (t_i, d_i) for the i-th patient, where t_i is the time to event or censoring, d_i is the indicator for event, 1 for observed, 0 for censored.

2.2 Blending hazard framework

Survival extrapolation is about predicting the unknown. Modellers have to ask themselves what is already known and how uncertain they are about the known. For optimal decision making, a good model may take various forms, but it has to be a framework that can incorporate all the information available and the associated uncertainty. In the context of modelling treatment effect waning, all the knowledge the company and the committee can have when a new treatment is appraised is:

- The short-term mortality rate for treatments in patients should perform the same as the mortality rate presented in the short-term trial;
- Treatments are likely to end up with approximately the same mortality rate in the long term due to treatment effect waning;
- The mortality rate in general population is very unlikely to change abruptly. The change from short-term mortality rate to long-term mortality rate should be a monotone and gradual process.

The rationale of the blending hazard method is to integrate all these knowledge together.

In the blended hazard method, the internal short-term data and the external long-term data are fitted by parametric models separately. Let $h_{int}(t|\theta_{int})$ and $h_{ext}(t|\theta_{ext})$ be the fitted hazard functions for internal data and external data respectively. The change in mortality rate over time can be expressed by the blended hazard function, a time-varying weighted average of internal hazard function and external hazard function

$$h_{blend}(t|\boldsymbol{\theta}) = [1 - \pi(t|t_1, t_2, a, b)] \times h_{int}(t|\theta_{int}) + \pi(t|t_1, t_2, a, b) \times h_{ext}(t|\theta_{ext})$$

where $\boldsymbol{\theta} = (\theta_{int}, \theta_{ext}, t_1, t_2, a, b)$ is the vector of model parameters, $\pi(t|t_1, t_2, a, b)$ is the weight function representing the rate of blending. The weight function is defined as the value of cumulative distribution function of Beta distribution $\text{Beta}(a, b)$ at $(t - t_1)/(t_2 - t_1)$

$$\pi(t|t_1, t_2, a, b) = \begin{cases} 0 \quad \text{for } 0 \leq t < t_1 \\ \text{F}_{\text{Beta}}\left(\frac{t-t_1}{t_2-t_1} \,\middle|\, a, b\right) \quad \text{for } t_1 \leq t < t_2 \\ 1 \quad \text{for } t \geq t_2 \end{cases}$$

where F_{Beta} is the cumulative distribution function of beta distribution $\text{Beta}(a, b)$, t_1 and t_2 are the starting time and ending time of blending respectively.

Before the blending interval, $t < t_1$, the weight of external data is 0, the blended hazard is identical to fitted internal hazard, survival extrapolation relies completely on information from the short-term trial. Within the blending interval, $t_1 < t < t_2$, the weight of external data increases from 0 to 1, the blended hazard gradually steers away from fitted internal hazard merges to fitted external hazard over time (Figure 5). As time goes on, the information from internal data is less reliable due to the evolve of disease and potential treatment effect waning. Information from external data becomes more reliable as the mortality rate is expected to eventually be the same as the external data after treatment effect vanishes in the long term. Since the uncertainty increases for fitted internal hazard and decreases for fitted external data, the weight function blends two sources of information to reach a reduced overall uncertainty in the blending interval. After the blending interval, $t > t_2$, the weight of external data is 1. As time far exceeds the short-term trial follow-up and there is no treatment effect, the extrapolation is purely informed by the external data.

Unlike some piecewise models designed for treatment waning where abrupt changes in hazard results in non-smooth survival, the blending hazard function

guarantees continuous change in hazard (the derivative of cumulative hazard function is continuous, cumulative hazard function is smooth), so the survival function is guaranteed to be a smooth function (Figure 6).

The blending hazard method is not merely a specific method of data fitting. It is a framework incorporating external data into trial-based extrapolation in a way that allows us to account for treatment waning. The framework provides the flexibility to be applied to a wide range of scenarios by adapting its 4 components: internal model, external model, blending interval and blending rate. Though the framework is flexible, there is no one-fit-for-all choice for these 4 components. Care must be taken to think about clinical and biological justification for these 4 components. The following sections provide a rough guidance on choosing these 4 components and explain implications of different choices of components.

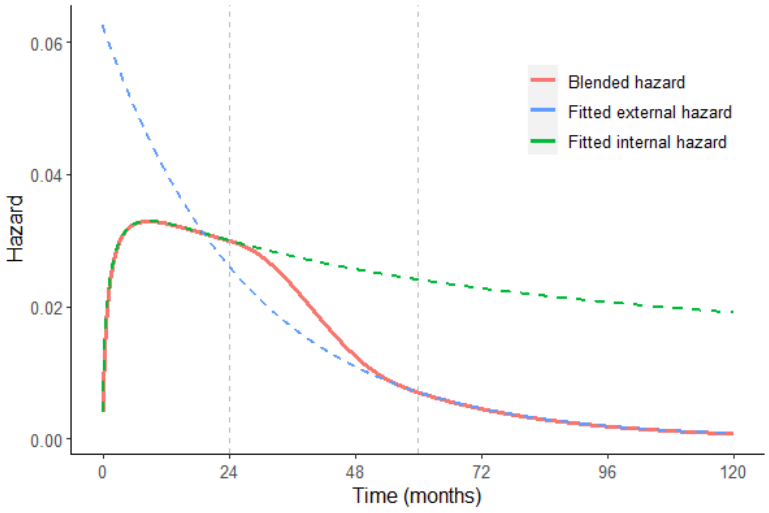

Figure 5 Graphical representation of the blending hazard method. The blending hazard methods enables hazard to gradually shift from internal fitted hazard to external fitted hazard during. The weight function controls the time and the rate of blending.

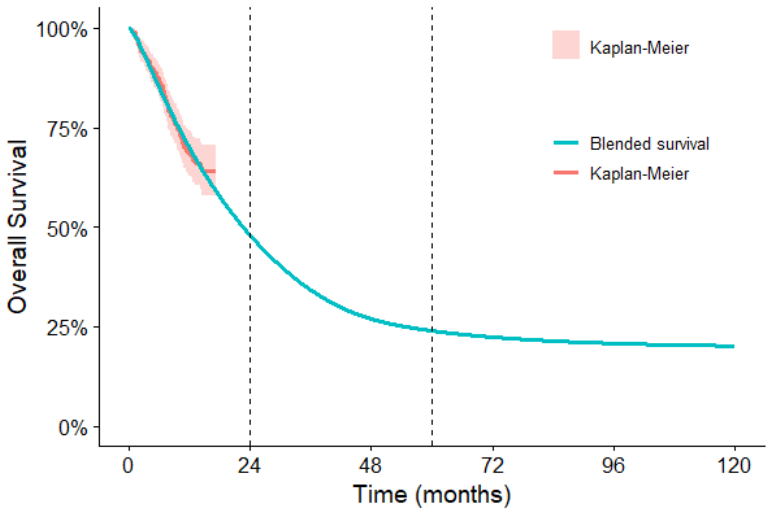

Figure 6 Graphical representation of survival curve from the blending hazard method. The survival function from blending hazard method is guaranteed to be a smooth function. There is no implausible change in the shape of survival function.

2.3 Modelling internal and external data

The internal data observed from the short-term trial provide accurate information on how survival behaves during the trial follow-up. The external data from other long-term studies provide information on how survival deemed to perform in the long term when there is no treatment effect. These two parts of information is used as inputs of the blending hazard method via fitted internal hazard and fitted external hazard. Since different models on the same data can generate different hazard functions for the same internal or external data, careful thought should be given to model selection when using the blending hazard method.

2.3.1 Modelling internal data

For internal data, conventional model selection for trial-based models lies in keeping the balance between providing good fit to the observed short-term data and generating plausible extrapolation in the long term: A flexible model that fits short-term data well may overfit observed data and lead to terrible extrapolation, and a model that generates reasonable extrapolation may not fit well to observed short-term data. However, within the blending hazard framework, the task of generating plausible long-term extrapolation is mainly left to the external data. The model selection should prefer relatively flexible models that provide good fit to the short-term data. Even if the fitted internal hazard is implausible in the long term, for example the extrapolated hazard is lower than general population mortality rate, the external data can still help to leverage this via the weight function.

The author suggests to start from basic standard parametric models and gradually increase the flexibility of model until there is no significant improvement in goodness of fit to the internal data. For each model, the hazard plots should be inspected to check if the model is flexible enough to capture the

shape of short-term hazard. AIC and BIC can be used to quantify goodness of fit to the short-term data and avoid unnecessarily complicated models. When models only have small difference in goodness of fit to internal data, the author suggests to visually compare their extrapolated hazard and survival in order to understand what they imply in the long term and decide which model is the most plausible. Although, as mentioned previously, the long-term extrapolation can be leveraged by the external data in the blending hazard framework, selecting relatively long-term plausible model when internal data are fitted well can help to reduce the "heavy-lifting" for the weight function.

2.3.2 Modelling external data

The external data are often selected to be mature enough to inform survival behaviours for both treatment arms in the long term. In the context of blending hazard method, the target of modelling external data is to obtain a model that accurately captures the long-term survival behaviours in the external data.

In practice, as external data usually show a generally decreasing trend in hazard, standard parametric models are often considered to be appropriate candidates. However, within the blending hazard framework, the author suggests to consider landmark models rebased at median follow-up of the short-term internal trial for modelling external data as a potential better option. This is because standard parametric models are fitted by optimising the fit to all observations, which implicitly involves a balance between fitting early observations well and fitting late observations well. However, the blending hazard method does not require the model for external data to provide good fit to early observations, as the blended hazard is dominated by fitted internal hazard before blending interval. i.e. fitted external hazard has no weight in initial period of time. Landmark models use the Kaplan-Meier for early observations and only use observations after landmark time point for model fitting, therefore they are better at characterising long-term survival. In terms of the choice of landmark time point, the chosen time point should be prior to the start of blending so that fitted external hazard is well defined within the blending interval. Since there are a relatively large number of patients before the median follow-up of the short-term internal trial, internal data can be considered to be relatively reliable before the median follow-up of the short-term internal trial and it is unlikely blend in external data before this time point. Therefore, median follow-up of the short-term internal trial can be considered to be a safe choice of landmark time point. Other choices of landmark time point can also be used as long as it is reasonable to believe the landmark time point is prior to the start of blending. Once the landmark time point is chosen, AIC, BIC and hazard plots can be employed as usual for choosing the landmark model with most appropriate parametric distribution.

2.4 The choice of weight function

The weight function is instrumental in the blending hazard method, as it decides when and how the information from internal data and external data are blended together. The fundamental distinction between the blending hazard method and other method for treatment effect waning also lies in the weight function:

Compared with the models with a piecewise construction (commonly used in appraisals), the weight function guarantees a continuous change in hazard function throughout the time horizon for the blending hazard method. Compared with models that rely on untestable assumptions (e.g. hazard ratio increase linearly, log hazard ratio increase linearly), the weight function in blending hazard method enables infinitely many scenarios of blending internal and external information.

The weight function $\pi(t|t_1, t_2, a, b)$ is specified by 4 parameters: t_1, t_2, a, b. Unlike parameters in fitted internal model and fitted external model θ_{int}, θ_{ext}, these parameters are not directly informed by data, but depend on people's subjective belief. However, this does not mean that the choice of these parameters is arbitrary. Careful thought about biological mechanism and sensitivity analysis still should be given to justify the choice and the impact of these parameters.

2.4.1 The choice of blending interval

The choice of blending interval defines the time when fitted internal hazard and fitted external hazard are combined.

The start of blending interval t_1 depends on the time when the fitted internal hazard would be no longer completely reliable. One option is to use the median follow-up of short-term trial. As the fitted hazard can be distorted by limited number of patients by the end of the short-term internal trial, modellers can be conservative about the shape of fitted internal hazard after the median follow-up, so that can be the time when external data start to have impact. Another option is to start blending at the time of treatment discontinuation. In this case, modellers are more confident with the fitted internal hazard and let it dominate longer. But after treatment discontinuation, it may not be reasonable to believe that the trend of fitted internal hazard will continue, thus the external data are combined to help account for the change in the trend of hazard.

The end of blending interval t_2 depends on the time when the two arms are believed to have the same hazard. i.e. hazard ratio is 1, there is no relative treatment effect. Since it is unknown how long the treatment effect will remain in the future, the choice of t_2 relies on treatment waning assumptions that come from expert opinion or past experience. For immuno-oncology therapies, a treatment effect that last for 3 to 5 years from the start of treatment has been considered plausible by the committee in a few past appraisals [30]. Hence, the author suggests to try different scenarios of t_2 between 3 years and 5 years in the blending hazard method for immuno-oncology therapies.

Since there is a large amount of uncertainty about blending interval, it is recommended to try multiple scenarios of blending interval and check if the blended hazard is plausible. Different blending intervals define different weight functions, hence imply different shape of blended hazard. Figure 7 demonstrates blended hazard functions for different blending intervals with the same Beta parameters for the same example data. It is always important to think about clinical plausibility of blended hazard. For example, in scenarios where blending interval is short, it is likely to observe a bump in blended hazard. Modellers can ask themselves whether it is plausible to have a bump in hazard during treatment waning. If yes, then think about how big the bump should be. In the end, if the resulting survival function is sensitive to the choice of blending

interval, then the most plausible blending interval can be used in the base case, other blending intervals should also be attached as alternative scenarios. If the survival function is not sensitive to the choice of blending interval, then all blending intervals can be presented as plausible scenarios.

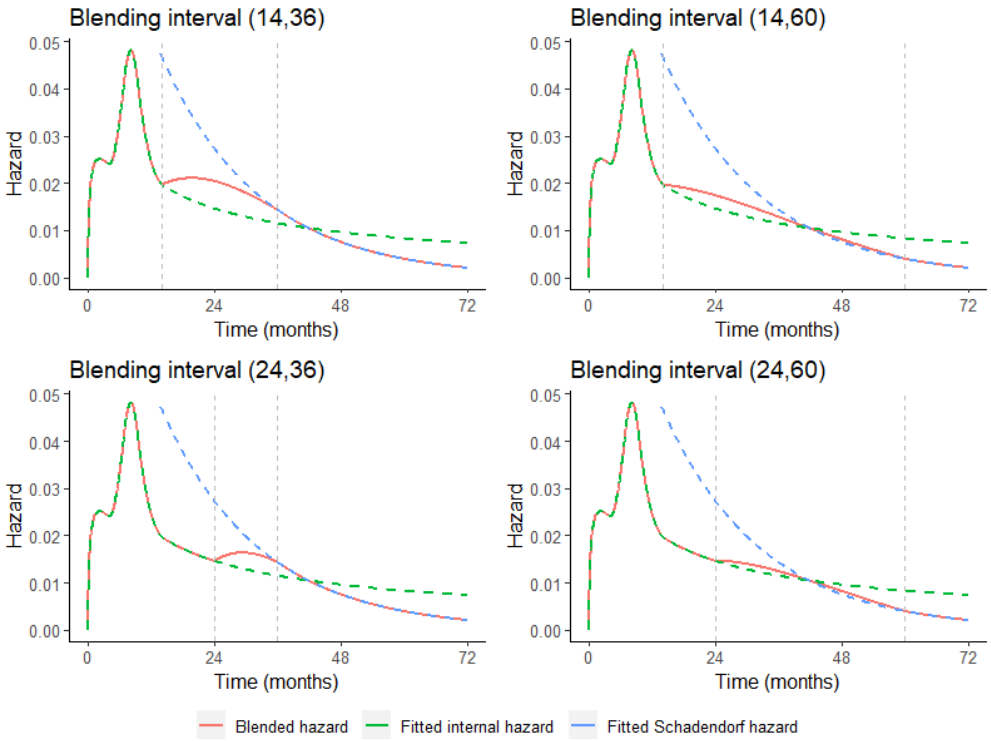

Figure 7 Blended hazard functions for different blending intervals with the same Beta parameters for the same example data. When the blending interval is relatively short, it is likely to observed a bump in the blended hazard and modellers should consider the clinical plausibility of the bump.

2.4.2 The choice of Beta parameters

Within the blending interval, the weight function is defined by the cumulative distribution function of a Beta distribution Beta(a, b). As shown in Figure 8, the shape of weight function varies with different settings of Beta parameters, so there could be many possible scenarios in the rate of blending internal information and external information. It is important to note that the rate of blending and the rate of waning are not the same thing: the rate of waning depends on the rate of blending and the gradients of fitted hazard functions. Because the fitted hazard functions are influenced by internal and external data, the rate of waning is partially data-driven. Whereas, the rate of blending is purely based on assumption about how the two pieces of information come together.

Though it is impossible to judge right or wrong for the choice, it is still worth the effort for modellers to understand implications behind these Beta parameters. Figure 9 describes blended hazard functions for different Beta parameters with the same blending interval for the same example data. When a is small, weight rise steeply right after t_1. As a gets larger, the initial increase in weight becomes more gradual, the blended hazard tends to follow fitted internal hazard longer at the

start of blending, representing increasing confidence with the internal data. When b is small, weight rise steeply right before t_2. As b gets larger, the final increase in weight becomes more gradual, the blended hazard tends to get closer to fitted external hazard at the end of blending, representing increasing confidence with the external data. When $a > b$, the blended hazard relies more on fitted internal hazard. When $a < b$, the blended hazard relies more on fitted external hazard. Just like choose the blending interval, modellers should try multiple scenarios of Beta parameters and consider whether these settings of Beta parameters result in plausible shape of hazard, especially for those blended hazard with a bump.

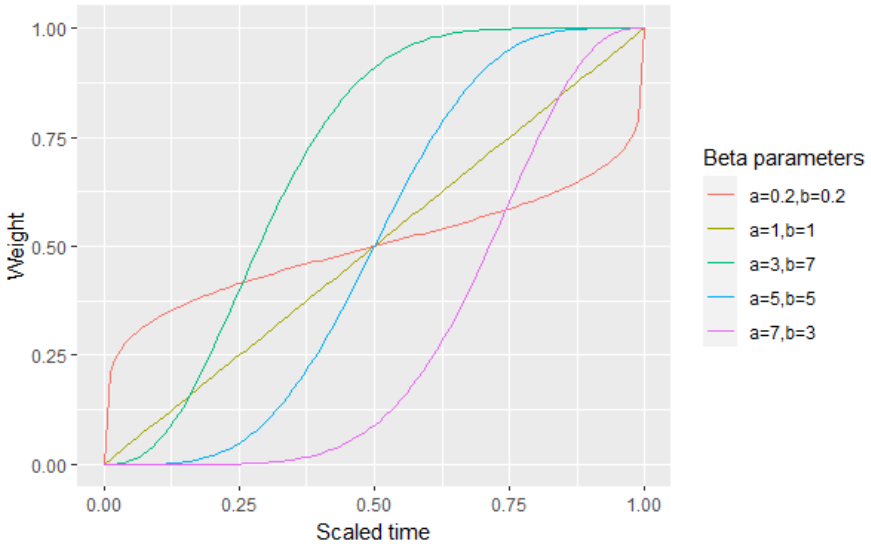

Figure 8 Shape of weight function varies with different values of Beta parameters. As a gets larger, the initial increase in weight becomes more gradual. As b gets larger, the final increase in weight becomes more gradual. When $a > b$, the weight stays low for a long time before a late increase. When $a < b$, the weight increases early before stays high for a long time.

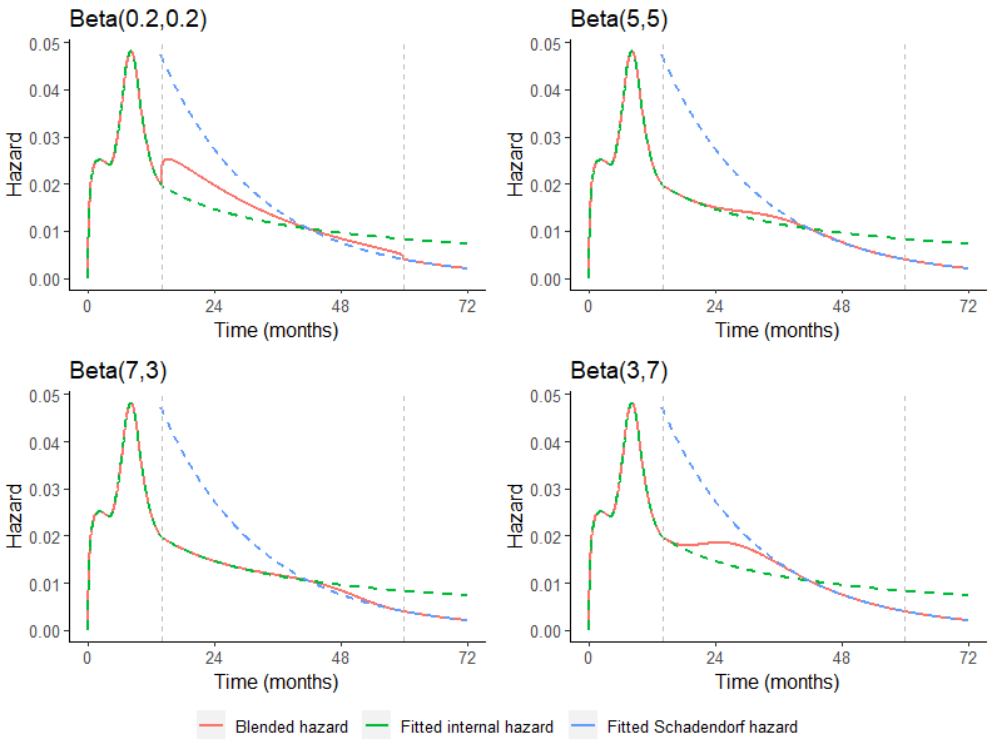

Figure 9 Blended hazard functions for different Beta parameters with the same blending interval for the same example data. As a gets larger, the blended hazard tends to get closer to fitted internal hazard after the start of blending. As b gets larger the blended hazard tends to get closer to fitted external hazard at the end of blending.

3 Results

3.1 Practical implementation on demonstrating case study

3.1.1 Data collection

The internal data of the demonstrating case study TA366 come from the KEYNOTE-006 trial that was used as the main trial to compare pembrolizumab with ipilimumab. The internal data are digitised from the OS Kaplan-Meier plot of KEYNOTE-006 trial in the TA366 committee paper [34].

For the external data, since both treatments are immunotherapies and the committee agreed there was no evidence suggesting the two treatment arms would have different rate of mortality in the long term, the best external data would be a long-term trial for ipilimumab on the same patient population. By reviewing available studies for ipilimumab on ipilumumab-naive patients with advanced melanoma, the treatment-naive subgroup in Schadendorf (2015) study [35] is considered to be a good source of external data for the following reasons: The Schadendorf study is a pooled analysis of long-term OS data from 10 phase II/III trials and 2 retrospective observational studies of ipilimumab. All the 12 studies target on unresectable or metastatic melanoma, which is in line with advance melanoma in KEYNOTE-006. 11 out of 12 studies include 3ug/mg dose, which matches ipilumumab dose in the KEYNOTE-006 trial. There is a presented subgroup analysis in the Schadendorf study with 7-year follow-up for ipilimumab-naive population, which matches the inclusion criteria of patients not previously treated with ipilumumab in the KEYNOTE-006 trial. The Schadendorf ipilimumab-naive population was considered the appraisal committee to be generalisable to the UK population. Considering the Schadendorf ipilimumab-naive population generally matches with the KEYNOTE-006 in the aspects listed above, the author believes it is not necessary to conduct population adjustments. Even if population adjustments are conducted, it is unlikely to exhibit a pronounced difference in external data. In addition, since the company did not perform any population adjustment in the original submission, using the same data as the company's submission is good for demonstrating how the extrapolation could have been improved if the blended hazard method were applied in the company's analysis. The external data are digitised from the OS Kaplan-Meier plot of ipilimumab-naive subgroup in the paper of Schadendorf et al. [35].

3.1.2 Modelling internal data

Modelling internal data for the demonstrating case study TA366 involves fitting two arms of KEYNOTE-006 trial OS data separately.

Pembrolizumab arm For pembrolizumab arm, standard parametric models are first fitted to the data. Since the smoothed hazard from data has an increasing then decreasing shape, generalised gamma, loglogistic and lognormal are chosen for further consideration (Figure 10). Comparing fitted hazard with smoothed hazard, none of these models can capture the decreasing trend in hazard before the end of trial (Figure 11). Although the shape of smoothed hazard can be

sensitive before the end of trial due to limited number of patients, the hazard before median follow-up is believed to be relatively reliable, and we expect the model to generally capture the trend of smoothed hazard before median follow-up. As fitted hazard from standard parametric models do not fulfil this, it suggests the standard parametric models are not flexible enough to provide a good fit to the data (black dashed lines represent median follow-up in Figure 11).

To allow more model flexibility, three kinds of restricted cubic spline models are considered: spline hazard, spline odds and spline normal. The 1-knot spline models have no big difference in hazard compared to previous standard parametric models, thus are omitted for brevity. The 2-knot spline models are better at capturing the decreasing trend, as the smoothed hazard falls into confidence intervals of fitted hazard before median follow-up. A potential problem is that the smoothed hazard are very close to margins of those confidence intervals and seems to overestimate the true hazard from 12 month to the median follow-up, which is likely to continue overestimate hazard in extrapolation (Figure 12). Therefore, model complexity continues to escalate. The 3-knot spline models provide very good fit to the data until the median follow-up, the difference between smoothed hazard and fitted hazard is minor (Figure 13). By calculating AIC and BIC for the models (Table 2), there is no significant increase in both AIC and BIC and AIC is slightly in favour of the escalation of flexibility from 2-knot spline to 3-knot spline. According to evidence above, the author believes 3-knot spline models are flexible enough and it is not necessary to further escalate model flexibility.

As for the choice of 3-knot spline hazard, 3-knot spline odds and 3-knot spline normal, their differences are negligible in the observed time, so the next step is to compare their extrapolated hazard and survival in order to understand what they imply in the long term. Figure 14 shows that 3-knot spline odds and 3-knot spline normal provide almost the same extrapolation in both hazard and survival scale, whereas 3-knot hazard model implies a higher hazard and a lower survival in the long-term comparing with the other two models. The author considers the 3-knot spline hazard model to be inferior as the large value in hazard and the fast decrease in survival at about 10 years for advanced melanoma are unusual, it is generally believed that hazard will be low and survival will reach a plateau at that time. Finally, the model for internal pembrolizumab data is chosen to be the 3-knot spline normal model due to its small advantage in AIC and BIC over the 3-knot spline odds model.

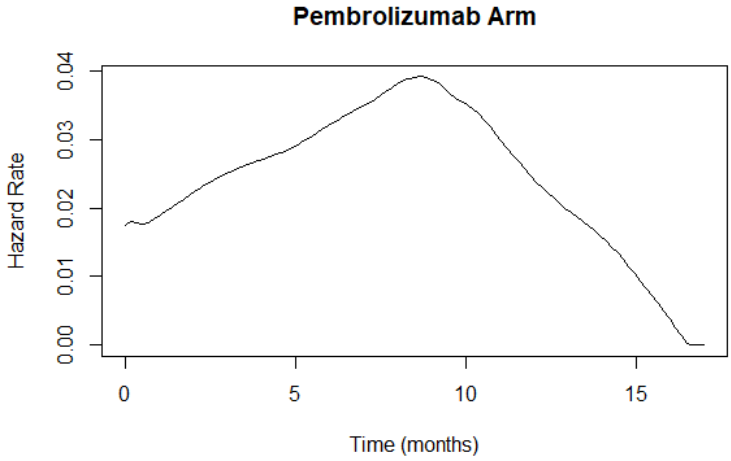

Figure 10 The smoothed short-term hazard for pembrolizumab arm has an increasing then decreasing shape. So Generalised Gamma, loglogistic and lognormal standard parametric models are considered for further investigation.

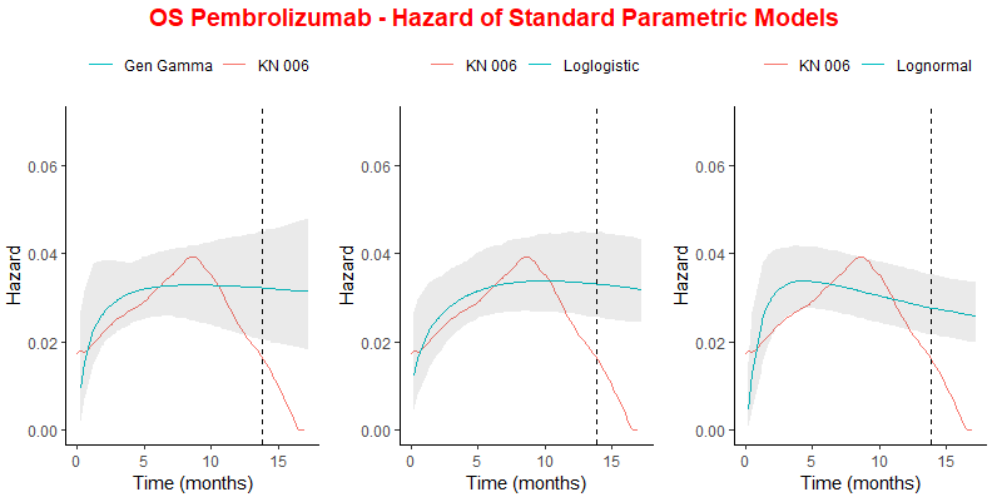

Figure 11 Hazard of standard parametric models for pembrolizumab arm. None of these standard parametric models can capture the decreasing trend in short-term smoothed hazard before the end of trial. KN 006 abbreviates for KEYNOTE-006.

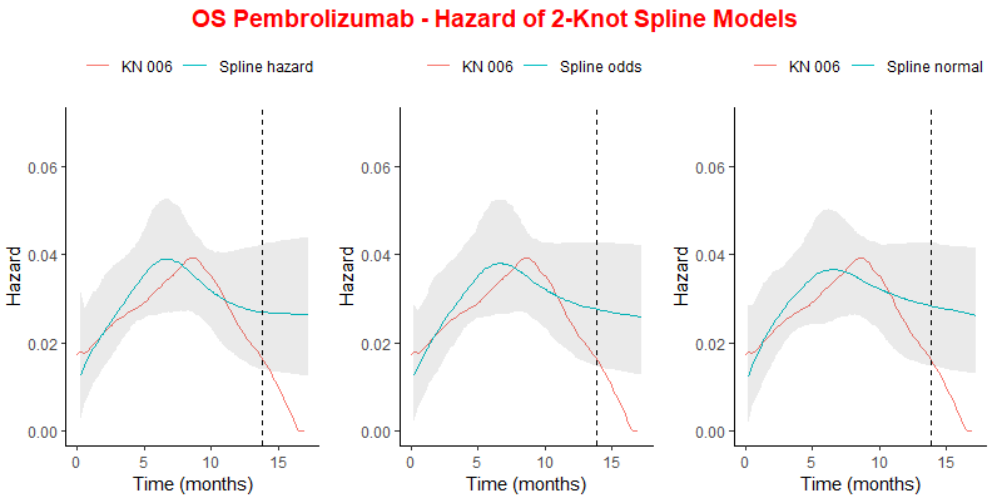

Figure 12 Hazard of 2-knot spline models for pembrolizumab arm. 2-knot spline models capture the decreasing better, but possibly overestimate hazard in the tail.

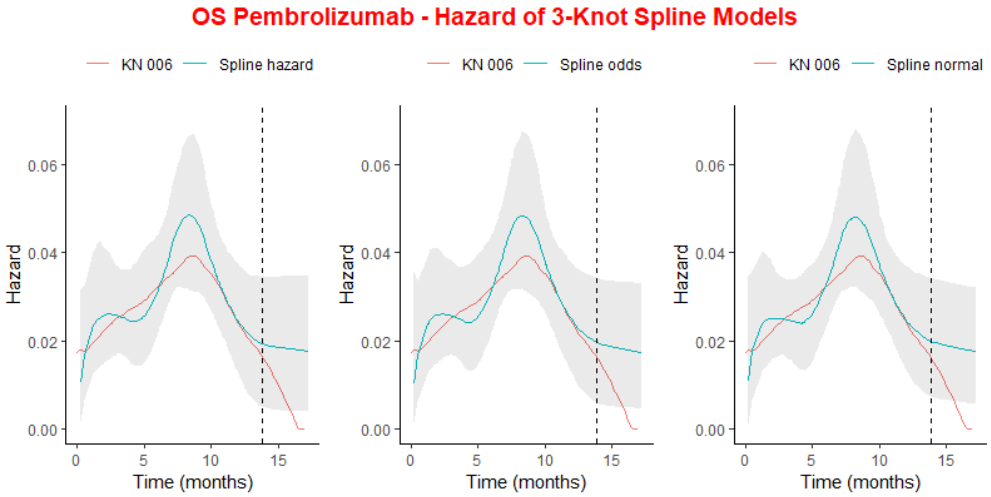

Figure 13 Hazard of 3-knot spline models for pembrolizumab arm. All 3-knot spline models provide very good fit to internal data.

Table 2: AIC and BIC of models in pembrolizumab arm.

Models	AIC	BIC
Gen Gamma	833.11	843.98
Loglogistic	830.89	838.14
Lognormal	832.43	839.67
2-knot spline hazard	833.84	848.34
2-knot spline odds	833.89	848.39
2-knot spline normal	833.95	848.44
3-knot spline hazard	831.83	849.95
3-knot spline odds	831.78	849.90
3-knot spline normal	**831.74**	**849.86**

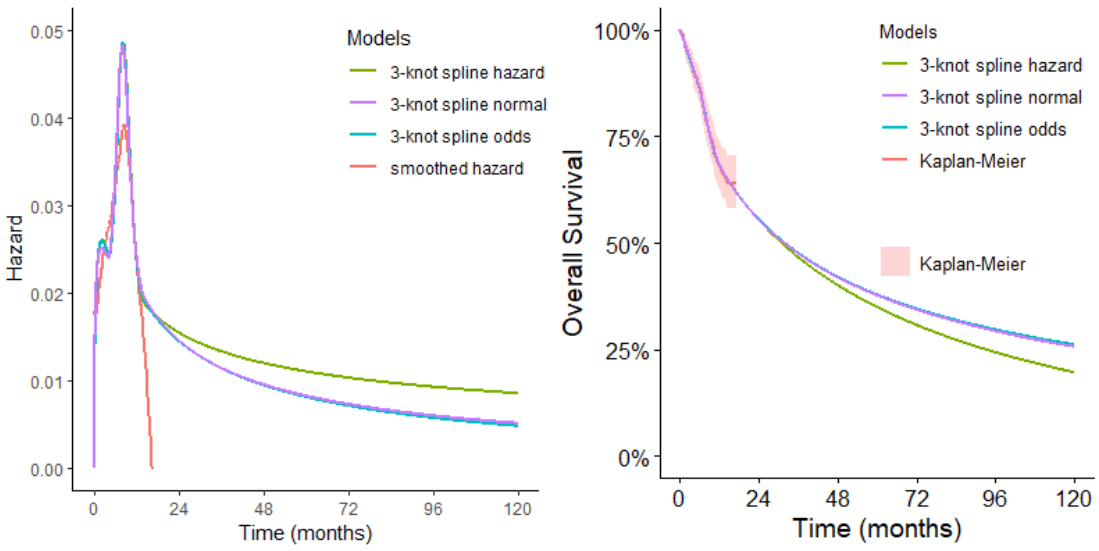

Figure 14 Extrapolated hazard and extrapolated survival of 3-knot spline models for pembrolizumab arm. 3-knot spline hazard model predicts higher hazard and lower survival than the other two models.

Ipilimumab arm The modelling of internal data for ipilumulab arm follows the same procedure. Standard parametric models are first considered. As the smoothed hazard from observed data has an increasing then decreasing shape in general (Figure 15), generalised gamma, loglogistic and lognormal models are chosen for further consideration. By comparing the three fitted hazard functions with the smoothed hazard function, the loglogistic model and the lognormal model fail to capture the decreasing trend in hazard in the second half of trial follow-up, while the generalised gamma model provides a very good fit to the data (Figure 16).

To decide whether to use the generalised gamma model or to escalate flexibility, the best way is to escalate a bit and check if the benefit worth introducing the extra flexibility. Therefore, spline hazard model, spline odds model and spline normal model with one knot are fitted to observed data. Although all 1-knot spline models provide good fit to the data, there is no significant improvement in the model fit by visual comparison (Figure 17). But the calculated AIC and BIC suggest the use of 1-knot spline models introduces no meaningful improvement in goodness of fit to internal data (Table 3). Therefore, the author believes that there is not enough benefit in favour of 1-knot spline models against generalised gamma model.

A further justification for this statement is given by looking at extrapolations. Figure 18 presents extrapolated hazard and survival implied behind generalised gamma and 1-knot spline models. The smoothed hazard and Kaplan-Meier from the external Schadendorf data are superimposed to hazard plot and survival plot respectively. Since the ipilimumab arm is expected to have the same mortality rate as the Schadendorf study, a plausible extrapolation should have similar long-term hazard as the Schadendorf study and have long-term survival that is basically parallel to the Schadendorf Kaplan-Meier. In this sense, extrapolated hazard and survival from the generalised gamma model align the best with the Schadendorf data, suggesting extrapolation from generalised gamma model is relatively more plausible. Because of all the statements listed above, the model

for internal ipilimumab data is chosen to be the generalised gamma model.

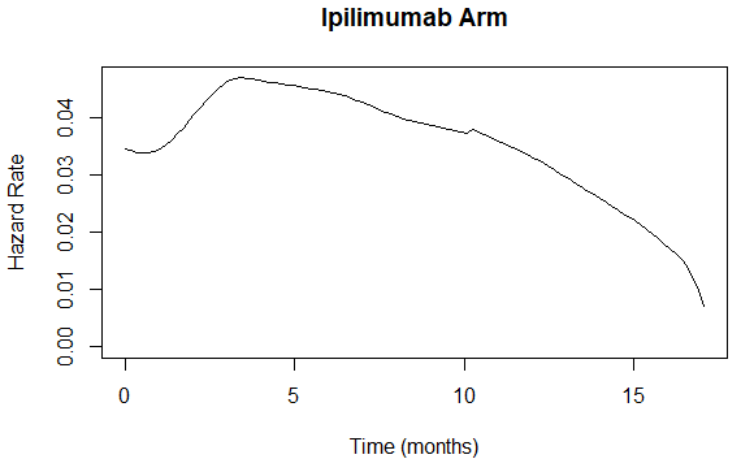

Figure 15 The smoothed short-term hazard for ipilimumab arm has an increasing then decreasing shape. So Weibull, Generalised Gamma, loglogistic and lognormal standard parametric models are considered.

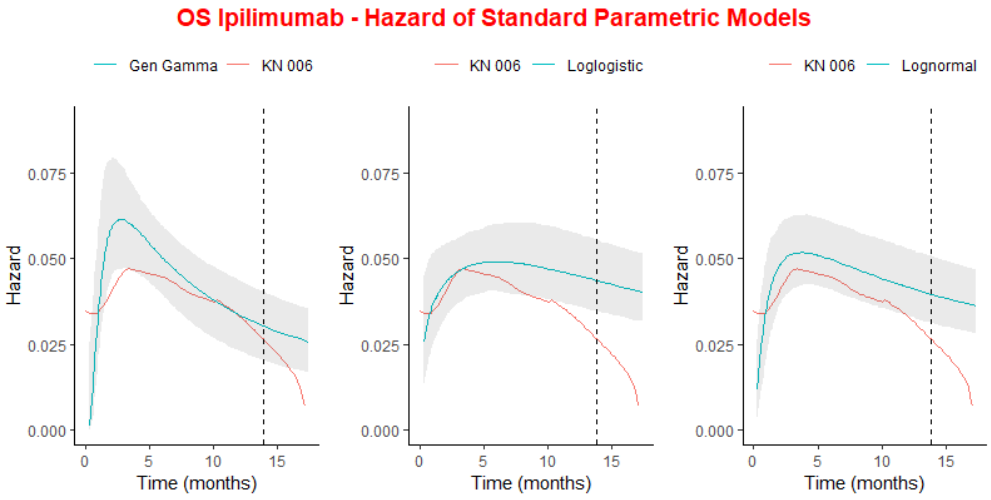

Figure 16 Hazard of standard parametric models for ipilumumab arm. The loglogistic model and the lognormal model fail to capture the decreasing trend in hazard in the second half of trial follow-up, while the generalised gamma model provides a very good fit to the data.

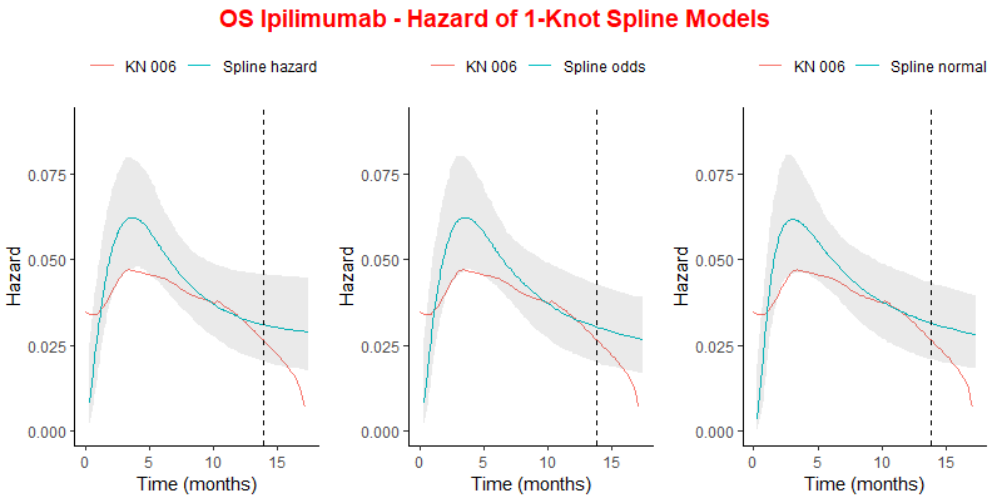

Figure 17 Hazard of 1-knot spline models for ipilimumab arm. There is no significant improvement in the model fit to internal data.

Table 3: AIC and BIC of models in ipilumumab arm.

Models	AIC	BIC
Gen Gamma	**922.87**	**933.75**
Loglogistic	931.27	938.52
Lognormal	924.79	932.05
1-knot spline hazard	925.00	935.88
1-knot spline odds	924.96	935.84
1-knot spline normal	923.52	934.40

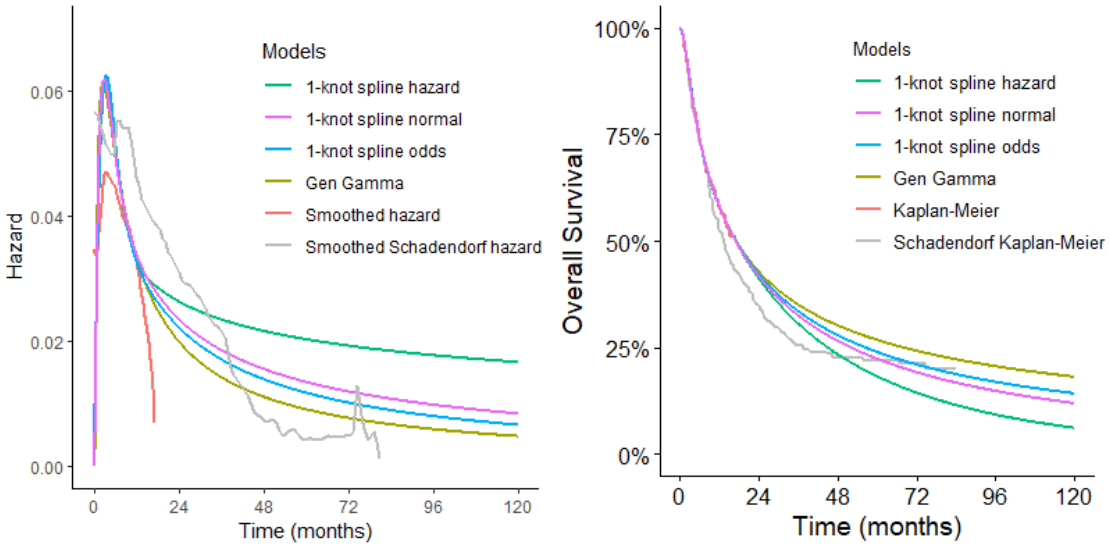

Figure 18 Extrapolated hazard and extrapolated survival of generalised gamma and 1-knot spline models for ipilimumab arm. Extrapolated hazard and survival from the generalised gamma model align the best with the Schadendorf data, suggesting extrapolation from generalised gamma model is more plausible.

3.1.3 Modelling external data

As stated earlier, the external data come from the ipilimumab-naive population in the Schadendorf study and have 7 years of follow-up. The median follow-up of KEYNOTE-006 trial is 13.85 months. Landmark models rebased at the median follow-up with different distributional forms are fitted to the Schadendorf data. As shown in Figure 19, the overall shape of smoothed Schadendorf hazard is decreasing, so Weibull, Gompertz, generalised gamma and loglogistic are potentially good candidates. The AIC and BIC of these landmark models are shown in Table 4, where both AIC and BIC indicate strong evidence against Weibull compared to the best candidate - Gompertz. So the landmark Weibull model is ruled out. By comparing fitted hazards from the rest of models with the smoothed Schadendorf hazard in Figure 20, the landmark Gompertz model is shown to be the best model in capturing the shape of smoothed Schadendorf hazard. The fitted survival curves shown in Figure 21 also validate the landmark Gompertz model provides the best fit to the Schadendorf Kaplan-Meier. Therefore, the model for the Schadendorf data is is chosen to be the landmark Gompertz model.

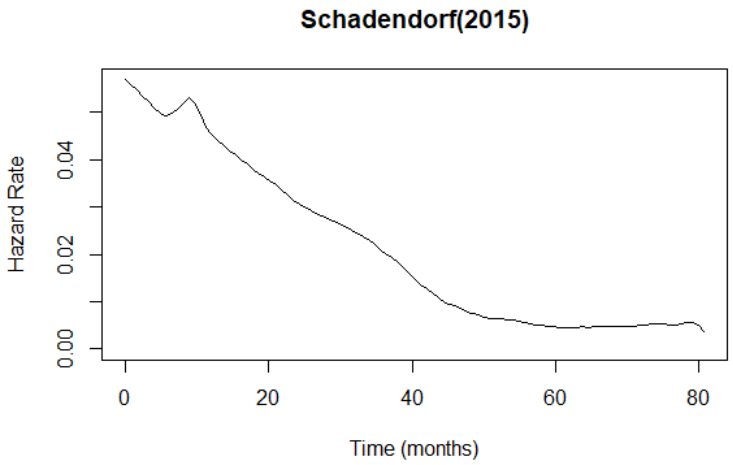

Figure 19 The smoothed hazard for the Schadendorf data has a generally decreasing shape. So Wiebull, Gompertz, Generalised Gamma and loglogistic standard parametric models are considered.

Table 4: AIC and BIC of rebased models for the Schadendorf data.

Models	AIC	BIC
Landmark Weibull	984.31	991.48
Landmark Gompertz	**971.93**	**979.09**
Landmark Gen Gamma	980.46	991.21
Landmark Loglogistic	977.88	985.05

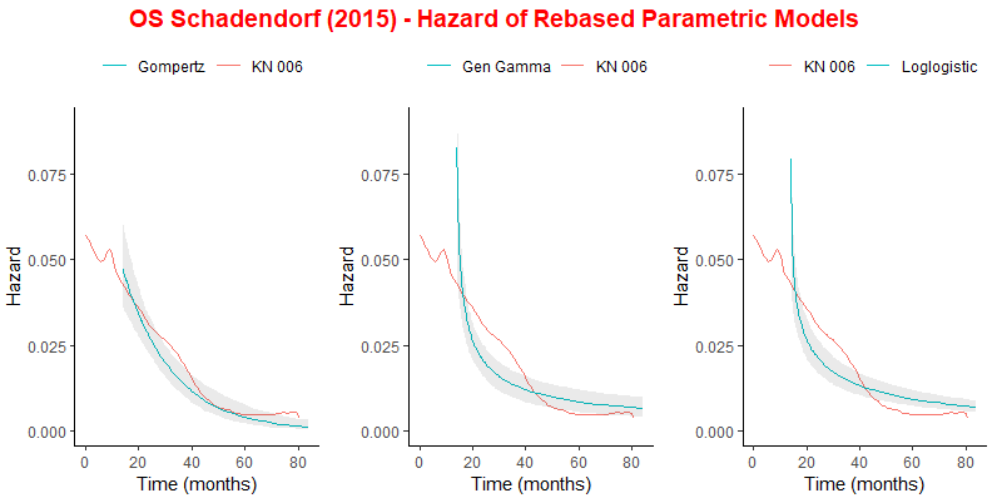

Figure 20 Hazard of landmark models for Scahdendorf data. The landmark Gompertz model captures the shape of smoothed Schadendorf hazard best.

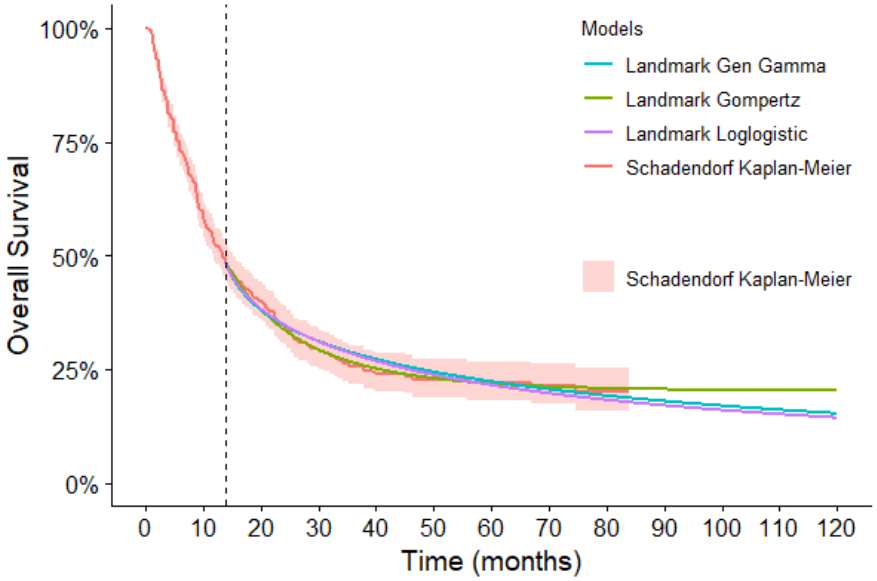

Figure 21 Survival of landmark models for Schadendorf data. The landmark Gompertz model provides the best fit to the Schadendorf Kaplan-Meier.

3.1.4 Choice of weight function

For the start of blending interval, one scenario is to use the median follow-up of the short-term KEYNOTE-006 trial, 13.85 months (rounded up to 14 for ease of computation), the other is to use the time of administrative discontinuation, 2 years (24 months), as the company proposed a 2-year stopping rule for the treatment. So 14 and 24 months are 2 scenarios considered for the start of blending interval. For the end of blending, Table 5 summaries treatment waning assumptions accepted by the committee in past pembrolizumab appraisals. As 5 out of 7 appraisals adopted treatment waning assumptions between 3 years and 5 years, it is reasonable to believe that a 3-to-5-year treatment waning is relatively clinically plausible for pembrolizumab. For this reason, 3 years and 5 years are

chosen to be 2 scenarios of the end of blending interval. Overall, 4 scenarios of blending interval are considered: (14 mo,36 mo), (14 mo,60 mo), (24 mo,36 mo), (24 mo,60 mo).

Table 5: Summary of accepted treatment waning assumptions in past pembrolizumab appraisals. (Selectively taken from Kamgar et al. paper [30])

TA	Waning assumption accepted by the committee
TA737	Gradual waning of hazard from year 5 to year 7
TA770	Equal hazard after 5 years
TA692	Equal hazard after 3 years
TA683	Gradual waning of hazard from year 2 to year 5
TA661	Equal hazard after 5 years
TA650	Equal hazard after 7 years
TA531	Equal hazard after 3 to 5 years

For the Beta parameters, it is suggested to provide scenarios with different shapes of weight function. In order to cover the shapes of weight function when (1) a, b are small, (2) a, b are large, (3) $a > b$ and (4) $a < b$, 4 scenarios of Beta parameters are considered: (0.2,0.2), (5,5), (3,7), (7,3).

In the end, with combinations of 4 scenarios of blending interval and 4 scenarios of Beta parameters, 16 scenarios of weight function are considered in total.

3.2 Extrapolated hazard and survival

This section shows the result of applying the blending hazard method for the TA366 case study. As explained in the previous section, the models for internal KEYNOTE-006 data are chosen to be a 3-knot spline normal model and a generalised gamma model for pembrolizumab arm and ipilumumab arm respectively. The model for external Schadendorf data is chosen to be a landmark Gompterz model rebased at KEYNOTE-006 median follow-up. For the weight function, 4 scenarios for blending interval ((14 mo,36 mo), (14 mo,60 mo), (24 mo,36 mo), (24 mo,60 mo)) and 4 scenarios for Beta parameters ((0.2,0.2), (5,5), (7,3), (3,7)) are considered, leading to 16 scenarios for the weight function and 16 corresponding models.

To assess the performance of these models, the updated 7-year follow-up data of the KEYNOTE-006 trial [38] are used to compare with the 16 scenarios of extrapolation from the blending hazard method. The piecewise extrapolation from the company's base case (KEYNOTE-006 OS Kaplan-Meier for the first 13 months and model fit to the Schadendorf data after 13 months) is also used to illustrate how much the extrapolation could have been improved if the blending hazard method were employed. To make this section concise, only the results from a selected model with weight function $\pi(t|t_1 = 14, t_2 = 60, a = 5, b = 5)$ are presented here, results from the other 15 models can be found in `https://github.com/JingqiZhu/TrtWaningSurvExtrap`.

Figure 22 demonstrates the blended hazard superimposed with the smoothed hazard from 7-year follow-up and the hazard in company's base case. The blended curve for pembrolizumab predicts hazard almost perfectly after 3 years. Specifically, the little bump in the smoothed hazard of pembrolizumab between 2

years and 3 years, which is hard to capture by conventional methods, can be captured via blending. The blended curve for ipilumumab predicts hazard very well before 2 years, but overestimates the hazard from 2 to 4 years and underestimates the hazard after 4 years. Since the ipilumumab arm is assumed to have the same hazard as Schadendorf study in the long term but the true long-term follow-up does not fully support this, it is possible that the patient characteristics in external Schadendorf study are biased and proper population adjustment should have been conducted to alleviate the difference in patient characteristics. Compared to the company's base case where the use of Schadendorf data from 13 months to 40 months largely overestimates the hazard, the hazard from blended method is much closer to the true smoothed hazard for both arms, indicating better extrapolations for both arms.

Figure 23 demonstrates the blended survival superimposed with the Kaplan-Meier from 7-year follow-up and the extrapolated survival in the company's base case. The blending hazard method gives extremely accurate prediction to long-term survival for both arms. On the other hand, the company's base case extrapolation is poor. The company's base case underestimates survival for both arms from 13 months due the the overestimation of hazard. This is especially true for pembrolizumab arm, where there is a pronounced discrepancy between company's extrapolated survival and the true long-term Kaplan-Meier.

This model with the specific setting of weight function $\pi(t|t_1 = 14, t_2 = 60, a = 5, b = 5)$ is not the only scenario that works well for the case study. Similar hazard plots and survival plots for the other 15 scenarios show that most of them provide good extrapolations that is basically in line with true 7-year follow-up and the extrapolation is not very sensitive to the choice of blending interval and Beta parameters. Compared to the company's base case, the blending method extrapolation is more consistent with the updated 7-year follow-up data than the company's base case extrapolation in all scenarios.

It is worth noting that the implementation of blending hazard method uses exactly the same data as the company's submission in 2015. No additional information that was not available in 2015 has been brought into the model. This can be served as a fair comparison to show that the blending hazard method is superior to the company's method.

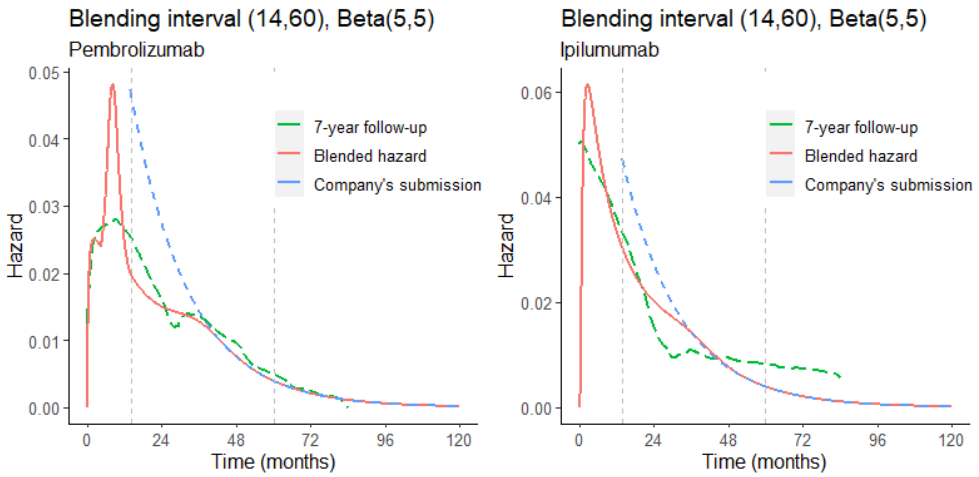

Figure 22 Blended hazard using blending interval (14,60) and Beta(5,5) vs smoothed hazard from 7-year follow-up and the hazard in company's base case.

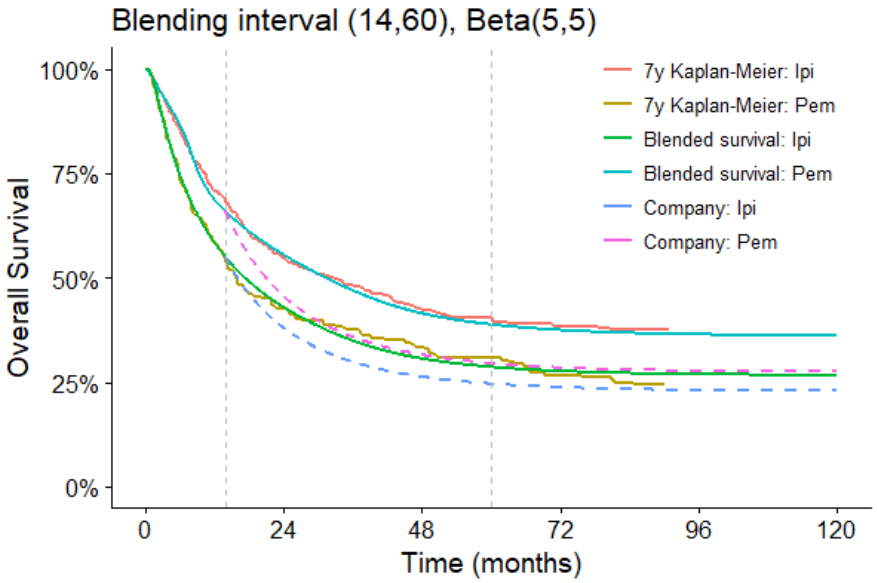

Figure 23 Blended survival using blending interval (14,60) and Beta(5,5) vs Kaplan-Meier from 7-year follow-up and the survival in company's base case.

3.3 Restricted mean survival time

In the context of health economic evaluation, survival extrapolation is used to calculate the restricted mean survival time (RMST) in the time horizon. The RMST estimates the life years (LY) of treated patients, which is directly used as an input for economic models. Apart from visual inspection of extrapolated hazard and survival, RMST offers another straightforward way to assess model results.

As the average age of patients in the KEYNOTE-006 trial is 59, to cover the maximum expected life expectancy, the calculation of RMST uses a time horizon of 30 years. This time horizon is also used in the company's original submission, so it allows us to make fair comparison between our method and the company's method. Table 6 shows the calculated RMST of two arms for the 16 scenarios of

blending hazard method. Table 7 shows the corresponding incremental RMST for the 16 scenarios of blending hazard method. The estimated RMST varies between 10.58 and 12.00 for pembrolizumab, varies between 8.43 and 9.01 for ipilumumab. The estimated incremental RMST spans from 2.10 to 3.08. Both RMST and incremental RMST are not sensitive to the choice of blending interval and Beta parameters.

The most ideal way to assess the estimates of RMST and incremental RMST is to compare with the true underlying RMST and incremental RMST. Certainly, the true underlying RMST and incremental RMST is unknown, but since we have access to the updated 7-year follow-up data now, the estimated RMST and incremental RMST estimated from a standard parametric model for the 7-year follow-up data can be used as surrogate of the true underlying RMST and incremental RMST. It is examined that a Gompertz model fit best to the 7-year follow-up data and the resulting estimated RMST and incremental RMST from the Gompertz model are shown in Table 8. In general, the blending hazard method tends to slightly underestimate RMST for pembrolizumab, but slightly overestimate RMST for ipilumumab, which results in underestimating the incremental RMST.

Compared with the company's base case, the blending hazard method gives much more accurate estimate for both RMST and incremental RMST. The company's base case underestimated a lot for all arms RMST as well as incremental RMST. This means that the company's base case undervalues the extending-life effect for both treatments as well as the incremental benefit of using pembrolizumab against ipilumumab. The company could have justified their new treatment to be more effective, hence more cost-effective, if the blending hazard method were employed. (The calculated RMST of company's base case is not the same as in the committee paper, because the company used Balch (2001) registry data [36] for extrapolation after 10 years. Since the blending hazard method only uses one source of external data, the use of registry data in company's base case is ignored in order to present a fair comparison between methods.)

Table 6: RMST for different scenarios of blending interval and Beta parameters

RMST (Pem/Ipi)		Beta parameters			
		(0.2,0.2)	(5,5)	(7,3)	(3,7)
Blending interval	(14,36)	10.77/8.53	11.18/8.75	11.61/8.93	10.58/8.48
	(14,60)	10.87/8.43	11.87/8.96	11.96/8.88	11.34/8.80
	(24,36)	11.59/8.91	11.68/8.96	11.83/9.01	11.49/8.89
	(24,60)	11.57/8.73	12.00/8.96	11.94/8.82	11.87/8.99

Table 7: Incremental life years gained for different scenarios of blending interval and Beta parameters.

Incremental RMST		Beta parameters			
		(0.2,0.2)	(5,5)	(7,3)	(3,7)
Blending interval	(14,36)	2.24	2.43	2.68	2.10
	(14,60)	2.44	2.91	3.08	2.54
	(24,36)	2.68	2.72	2.82	2.60
	(24,60)	2.84	3.04	3.12	2.88

Table 8: RMST and incremental RMST of company's base case and Gompertz model for 7y follow-up.

Model	RMST (Pem/Ipi)	Incremental RMST
Gompertz model for 7y follow-up	12.01/8.12	3.89
Company's base case	8.39/7.00	1.39

4 Discussion

4.1 Summary of the study

There is a growing demand for modelling treatment effect waning in survival extrapolation with the development of immuno-oncology. To inform credible cost-effectiveness analysis for policy makers, it is essential to overcome the challenge of survival extrapolation with the presence of treatment effect waning. However, the modelling method targeting treatment effect waning is still underdeveloped: There is no guidance from NICE TSD, only a few researches focus on treatment effect waning and existed methods rely on untestable or even implausible assumptions on hazard ratio. In this dissertation, the author has proposed the blending hazard method as a possible solution to bridge the gap.

The blending hazard method fits separate models to internal short-term data and external long-term data and combine the two fitted hazard functions using a time-varying weight function for each treatment arm. The method enables the extrapolated hazard to be initially dominated by the fitted internal hazard, then gradually approaches the fitted external hazard over a blending interval, and finally be dominated by the fitted external hazard. During the blending interval, the weight function is defined to be a cumulative distribution function of a Beta distribution and it controls the rate of blending internal and external information.

A model from the blending hazard method is specified by 4 components: model for internal data, model for external data, blending interval and Beta parameters. This dissertation provides a rough guidance for choosing these components. With the existence of external data, the first priority for internal models is to fit internal data well. The internal models should be selected by flexibility escalation - escalating flexibility gradually until it provides reasonably good fit to internal data. The external model should be a model that provides good fit to long-term hazard in external data. It is suggested to consider a landmark model rebased at median follow-up of the short-term trial in order to maximise the fit to external data in the long-term. The start of blending interval should chosen to be the time when extrapolation from internal model is not fully convincing. The end of blending interval should chosen to be the time when treatment effect is deemed to vanish. The choice of Beta parameters should reflect modellers' confidence in fitted internal hazard and fitted external hazard. The more modellers are confident in fitted internal hazard, the larger the first Beta parameter should be. The more modellers are confident in fitted external hazard, the larger the second Beta parameter should be. Since the choice of blending interval and Beta parameters are main based on assumptions, modellers should provide multiple scenarios of them so that their impact on final result can be well understood.

In this dissertation, TA366 is used as a case study to demonstrate the blending hazard method. Results from 16 different scenarios of blending interval and Beta parameters are compared to the updated 7-year follow-up. Extrapolation and restricted mean survival time are consistent with the updated data for most scenarios. Comparing to the company's base case method, with exactly the same data all scenarios generate more accurate extrapolation.

4.2 Strengths of the method

The author is convinced that a good extrapolation method should have 3 main characteristics:

- Make full use all of the knowledge available
- Account for the uncertainty about those knowledge
- Easy to implement

For the task of modelling treatment effect waning, only 3 things are known: (1) short-term hazard is the same as internal data; (2) long-term hazard will be the same as external data; (3) treatment effect is likely to wane from the time of discontinuation until 3-5 years, which can lead to unusual shape in hazard, but hazard will not change suddenly. The piecewise model in the company's base case fails to utilise the information from (3), so results in a implausible sudden change in hazard and poor extrapolation. Even if the company were uncertain about the 3-5 years range, assuming 2-10 years should be plausible to use. The blending hazard method utilises all of these knowledge and formulated them into internal model, external model and weight function respectively.

The advantage of the blending hazard method is not only reflected in using all sources of knowledge, but also in making the full use of these knowledge. Unlike in the piecewise method, in the blending hazard method, internal information is not used in internal model exclusively and external information is not used in external model exclusively. Instead, information flows freely among internal model, external model and blending process as shown in Figure 24: When fitting internal model, external data are borrowed to help select models that give plausible extrapolation; When fitting external model, internal information about median follow-up of short-term trial is borrowed to help decide the landmark time point; When blending, both internal and external fitted hazard are used, which means both internal information and external information go into the blending process. The blending hazard method has a more connected information propagation network, thus utilises available information better.

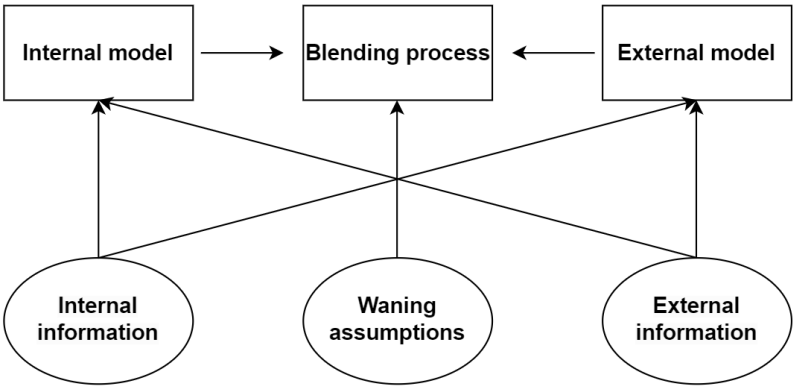

Figure 24 Information propagation network of the blending hazard method.

Modelling treatment waning involves a large amount of uncertainty associated with the time and the rate of waning. Although, the piecewise method allows

different scenarios of cutting time point, it completely ignores the uncertainty associated with the rate of waning. The sudden change in hazard is an assumption that is too strong to characterise the limited knowledge on waning. The linear waning method allows linear change in hazard ratio, which is an untestable and possibly implausible assumption. Since there is no information on how the hazard will change, a linear waning assumption implicitly ignores a large amount of uncertainty. In the blending hazard method, different scenarios of blending interval account for the uncertainty about waning time, different scenarios of Beta parameter account for the uncertainty about the blending rate. The uncertainty about blending rate and the uncertainty within internal and external fitted models together constitute the uncertainty about waning rate. When people have better understanding on the treatment waning mechanism, uncertainty associated with wight function can be updated.

The blending hazard method is also easy to implement. The modelled hazard takes a simple form of the weight average of fitted internal hazard and fitted external hazard. The primary uncertainty about blending is characterised by manipulating different scenarios of weight function. Compared to some Bayesian methods, where prior elicitation, computational software (BUGS, stan) and coding experts are needed [39], this simplicity of this method implies it can be more easily accepted by HTA companies. (Though the idea is Bayesian, frequentist way of implementation reduces companies' workload.)

4.3 Limitations and extensions of the method

The blended hazard function is generally not sensitive to the choice of weight function, but can be relatively easily influenced by the choices of internal and external models. As the blended hazard is a weighted average, it can only lie between the fitted internal hazard and fitted external hazard. When the models are not selected appropriately, it is possible that there would not much area between the two fitted hazard and the limited area is not enough to characterise a hazard that is close to the underlying true hazard. In this case, the extrapolation would deviate from the truth no matter how we chose parameters for the weight function.

When selecting internal model for the case study TA366, the flexibility escalation resorts to adding knots to spline models. Other flexibility escalation schemes, for example using M-spline or mixture cure models, are also potentially useful and can be considered in future research.

The blending hazard method is shown to be useful on the case study TA366. The method can also be applied to other NICE technical appraisals on different immunotherapies for different indications to further evaluate its extrapolation performance.

In practical survival extrapolation , it is sometimes required to use more than one sources of external data to inform long-term extrapolation. The blending hazard method shown here only demonstrate the case of blending internal data with one source of external data (blending two components). Attempting to blend multiple sources of external data would be a useful area for further research.

5 Conclusions

The dissertation proposes the blending hazard method as a solution for modelling treatment effect waning in survival extrapolation. The method fits separate models to internal short-term data and external long-term data and combine the two fitted hazard functions using a time-varying weight function for each treatment arm. The extrapolated hazard is initially dominated by the fitted internal hazard, then gradually approaches the fitted external hazard over a blending interval, and is finally dominated by the fitted external hazard. The weight function, characterised by a cumulative distribution function of a Beta distribution, controls the rate of combining internal and external information during the blending interval.

The NICE TA366 has been used as a case study to illustrate practical implementation of the method. The corresponding results are shown to have a good consistence with the updated 7-year follow-up and perform a great improvement from the original company's base case.

On the basis of this dissertation, the author recommends to consider using other flexibility escalation schemes for internal model or blending more than one sources of external information for future area of research. The author also calls for more methods to be developed for modelling treatment effect waning in survival extrapolation.

References

[1] Gianluca Baio. *Bayesian methods in health economics*. London: CRC Press, 2012.

[2] Elisabeth Fenwick and Sarah Byford. "A guide to cost-effectiveness acceptability curves". In: *The British Journal of Psychiatry* 187.2 (2005), pp. 106–108.

[3] Gianluca Baio. "survHE: Survival Analysis for Health Economic Evaluation and Cost-Effectiveness Modeling". In: *Journal of Statistical Software* 95.14 (2020), pp. 1–47.

[4] David Lebwohl et al. "Progression-Free Survival: Gaining on Overall Survival as a Gold Standard and AcceleratingDrug Development". In: *The Cancer Journal* 15.5 (2009), pp. 386–394.

[5] Simone Amabile et al. "Clinical significance of distant metastasis-free survival (DMFS) in melanoma: a narrative review from adjuvant clinical trials". In: *Journal of clinical medicine* 10.23 (2021), p. 5475.

[6] Suad AK Shamis, Donald C McMillan, and Joanne Edwards. "The relationship between hypoxia-inducible factor 1α (HIF-1α) and patient survival in breast cancer: Systematic review and meta-analysis". In: *Critical reviews in oncology/hematology* 159 (2021), p. 103231.

[7] Sally Anne Garnett et al. "Comparing duration of response and duration of clinical benefit between fulvestrant treatment groups in the CONFIRM trial: application of new methodology". In: *Breast cancer research and treatment* 138 (2013), pp. 149–155.

[8] Daniel Gallacher, Peter Auguste, and Martin Connock. "How Do Pharmaceutical Companies Model Survival of Cancer Patients? A Review of NICE Single Technology Appraisals in 2017". In: *International Journal of Technology Assessment in Health Care* 35.2 (2019), pp. 160–167.

[9] National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal 2013*. `https://www.nice.org.uk/process/pmg9`. [Accessed: 12 August 2023]. 2013.

[10] Canadian Agency for Drugs and Technologies in Health. *Guidelines for the economic evaluation of health technologies: Canada*. 2017.

[11] Helen Bell Gorrod et al. "A review of survival analysis methods used in NICE technology appraisals of cancer treatments: consistency, limitations, and areas for improvement". In: *Medical Decision Making* 39.8 (2019), pp. 899–909.

[12] Nicholas Latimer. "NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data". In: *Report by the Decision Support Unit* (2011).

[13] Mark J Rutherford et al. "NICE DSU technical support document 21: Flexible methods for survival analysis". In: *Report by the Decision Support Unit* (2020), pp. 1–97.

[14] Patrick Royston and Mahesh K. B. Parmar. "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". In: *Statistics in Medicine* 21.15 (2002), pp. 2175–2197.

[15] Jodi Gray et al. "Extrapolation of survival curves using standard parametric models and flexible parametric spline models: comparisons in large registry cohorts with advanced cancer". In: *Medical Decision Making* 41.2 (2021), pp. 179–193.

[16] Benjamin Kearns et al. "Comparing current and emerging practice models for the extrapolation of survival data: a simulation study and case-study". In: *BMC medical research methodology* 21.1 (2021), pp. 1–11.

[17] Therese ML Andersson et al. "Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models". In: *BMC medical research methodology* 11.1 (2011), pp. 1–11.

[18] Nicholas R Latimer and Amanda I Adler. "Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness". In: *BMJ Medicine* 1.1 (2022).

[19] Christopher Jackson et al. "Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods". In: *Medical Decision Making* 37.4 (2017), pp. 377–390.

[20] Patrick Royston, M Parmar, and Douglas G Altman. "External validation and updating of a prognostic survival model". In: *Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London* 222 (2010).

[21] National Institute for Health and Care Excellence. *TA135: Pemetrexed for the treatment of malignant pleural mesothelioma*. `https://www.nice.org.uk/guidance/ta135`. [Accessed: 12 August 2023]. 2008.

[22] National Institute for Health and Care Excellence. *TA357: Pembrolizumab for treating advanced melanoma after disease progression with ipilimumab*. `https://www.nice.org.uk/guidance/ta357`. [Accessed: 12 August 2023]. 2015.

[23] James Larkin et al. "The predicted impact of ipilimumab usage on survival in previously treated advanced or metastatic melanoma in the UK". In: *PLOS ONE* 10.12 (2015), pp. 1–11.

[24] Fanni Soikkeli et al. "Extrapolating survival data using historical trial–based a priori distributions". In: *Value in Health* 22.9 (2019), pp. 1012–1017.

[25] Michael J. Sweeting et al. "Survival Extrapolation Incorporating General Population Mortality Using Excess Hazard and Cure Models: A Tutorial". In: *Medical Decision Making* 43.6 (2023), pp. 737–748.

[26] Zhaojing Che, Nathan Green, and Gianluca Baio. "Blended Survival Curves: A New Approach to Extrapolation for Time-to-Event Outcomes from Clinical Trials in Health Technology Assessment". In: *Medical Decision Making* 43.3 (2023), pp. 299–310.

[27] Federico Felizzi et al. "Mixture cure models in oncology: a tutorial and practical guidance". In: *PharmacoEconomics-Open* 5 (2021), pp. 143–155.

[28] Mario JNM Ouwens et al. "Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations". In: *Pharmacoeconomics* 37 (2019), pp. 1129–1138.

[29] Tai-Tsang Chen. "Statistical issues and challenges in immuno-oncology". In: *Journal for immunotherapy of cancer* 1 (2013), pp. 1–9.

[30] F Kamgar et al. "A review of treatment effect waning methods for immuno-oncology therapies in National Institute for Health and Care Excellence Technology Appraisals". In: *Value in Health* 25.12 (2022), S98.

[31] J Horscroft et al. "A review of differences in decision-making across NICE health technology assessments of nivolumab". In: *Value in Health* 22 (2019), S503.

[32] Patricia Guyot et al. "Extrapolation of Survival Curves from Cancer Trials Using External Information". In: *Medical Decision Making* 37.4 (2017), pp. 353–366.

[33] Christopher Jackson. "survextrap: a package for flexible and transparent survival extrapolation". In: *arXiv preprint arXiv:2306.03957* (2023).

[34] National Institute for Health and Care Excellence. *TA366: Pembrolizumab for advanced melanoma not previously treated with ipilimumab committee papers.* `https://www.nice.org.uk/guidance/ta366/documents/committee-papers`. Accessed: 15 August 2023. 2015.

[35] Dirk Schadendorf et al. "Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma". In: *Journal of clinical oncology* 33.17 (2015), p. 1889.

[36] Charles M Balch et al. "Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma". In: *Journal of clinical Oncology* 19.16 (2001), pp. 3635–3648.

[37] Patricia Guyot et al. "Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves". In: *BMC medical research methodology* 12 (2012), pp. 1–13.

[38] Caroline Robert et al. "Seven-Year Follow-Up of the Phase III KEYNOTE-006 Study: Pembrolizumab Versus Ipilimumab in Advanced Melanoma". In: *Journal of Clinical Oncology* (2023), JCO–22.

[39] Ash Bullement et al. "A Systematic Review of Methods to Incorporate External Evidence into Trial-Based Survival Extrapolations for Health Technology Assessment". In: *Medical Decision Making* 43.5 (2023), pp. 610–620.