D2.1 - A review of parametric model choice used to extrapolate RCT outcomes.

Caro, J., McGuire, A. & Wood, M.

Background

Many policy decisions in health economics are embedded in health technology assessments (HTAs). However, HTAs are often constrained with regards to the data upon which these assessments can be made. Whilst some of this data is available from clinical trials or real-world evidence, extrapolation over lifetime is generally necessary to capture the full treatment benefit of an intervention. This is because the clinical trial period may not cover the entire time to event or may not be generalisable to the broader population due to differing population characteristics. HTAs are particularly interested in final outcomes, survival. Moreover, the emergence of immunotherapies with their potential for cure has introduced additional complexities in making these extrapolations.

However, mortality is only one of many endpoints that they might consider. Other endpoints include quality of life adjusted survival (QALYs), disease progression and treatment discontinuation. In this paper we focus on the modelling of survival itself, but note aspects of extrapolating other outcomes where necessary.

There are several methods available for such extrapolation. These different methods can result in different estimates. As such, failure to systematically chose a model for extrapolation can create inconsistency and between HTAs. It is essential that a systematic approach is used so to demonstrate that extrapolation has been undertaken appropriately and so that decision makers can be confident in the results of the associated economic analysis. Consequently, inefficient decisions can be avoided.

Objective

The aim of the paper is to improve understanding of the current approach being used to guide model choice for extrapolation, and to critique existing frameworks for model choice, to improve the utilisation of such techniques within standard economic evaluation applications. This includes a discussion of the use of data from different sources, i.e. matching of data from randomised control trials with observational or Registry data banks. This is a relatively underutilised method to help guide model choice and development. We will outline the process and advantages of using external data to assist with model choice.

Methods

To better understand the current approach to model selection we first provide a background on the most used methods for the extrapolation of benefits and the techniques available to guide choice. To outline the approach used in practice, we review the survival analyses (methods used, and justification for these methods) included in National Institute for Clinical Excellence (NICE) Technology Appraisals between 2015-16 and 2019. We review NICE appraisals, as NICE is one of the earliest Health Technology Agencies to focus on the use of cost-effectiveness in assessing new Health Technology Appraisals (HTAs) and has developed

a clear methodological stance in the assessment of evidence. NICE methods are also utilised as a benchmark by other HTAs internationally. The NICE Decision Support Unit (DSU) provides technical support documents (http://nicedsu.org.uk/technical-support-documents/technical-support-documents/) to assist with HTA. NICE appraisals are also useful as WP2 uses members of NICE to evaluate the DICE model approach being developed within the WP.

We focused on HTAs of treatments for metastatic cancer and on the use of extrapolation within Markov models (a frequently used model in HTA where future states depend only on the current state) to maximise comparability whilst maintaining a feasible sample of HTAs. We are particularly interested in the applicability of extrapolation methods for use in the DICE model captured within WP2. Metastatic cancer is chosen partly because of the recent debate over the use of overall survival versus progression free survival as an endpoint for both trials in this area and extrapolated models2. As such, this provides us the opportunity to also discuss the decision process for choosing appropriate end points.

Results

We identified 47 relevant HTAs between 2015-16 and 2019. All but two of these HTAs provided some details of the decision-making process to choose a model to extrapolate survival benefits. Statistical tests were used in all but three HTAs (94%) and visual inspections were used in all but nine HTAs (80%). The plausibility of the distribution was a factor in model choice in 28 (60%) of HTAs, and expert opinion was used in 19 (40%). The use of external data is relatively limited being used in only ten (21%) of HTAs.

Several frameworks, guidelines or algorithms have been published to assist with model choice and help ensure models are chosen systematically. The most relevant of these is Latimer 2013, a framework produced by the Decision Support Unit of NICE. This framework has been used by a number of recent HTA submissions. We review this approach and a recent critique that focusses on data availability, which is a general problem encountered in this area.

Limitations

Our review of model justification in HTAs is reliant on information provided in company submissions. If there were factors that affected model choice which were not sufficiently (or explicitly) detailed in the HTA submission, and this occurs mainly because of commercially sensitive issues, they are not included in this review.

This paper attempts to describe the most used methods in HTAs, with a view to discussing the criteria to choose between them. As such this not a comprehensive discussion of survival analysis but aims to improve HTA decision making and the identification of appropriate criteria for use in defining health benefits as incorporated within the cost-effectiveness approach adopted by many HTAs.

Conclusion

The number of HTAs which have justified the choice of model for extrapolation has improved significantly over time. A previous review of HTAs submitted to NICE between 2000 and 2009 found very few HTAs justified model choice. The significant increase in model justification may reflect the

release of NICE guidelines (Latimer 2013). The framework presented in Latimer 2013 is relatively comprehensive and should be utilised in HTA.

However, there remains opportunity for improvement. Expert opinion and clinical plausibility should be used more frequently. In several cases the lack of consideration for clinical plausibility in company submissions was a reason for disagreement between the NICE and the company submission. The use of external data also offers opportunity to improve extrapolation and model choice. (https://www.iwmf.com/sites/default/files/docs/publications/PFS_QOL_Dec2018.pdf). It's unclear what drives the low usage of external data, it could be access to the required data, or awareness/capability of the approach. This is an area worthy of additional consideration.

Transparency in detailing the justification behind model choice in the extrapolation of survival benefits in HTA could be improved. The justification for choosing a model, and the limitations of chosen models should always be explicitly made. It is not sufficient to say a model was chosen as it was considered to have the best fit, without detailing the criteria which defined best fit. The use of quality of life adjustments and extrapolation remains rudimentary.

1. Introduction

Health technology assessments (HTAs) are fundamentally interested in the provision of value for money associated with the introduction of new health care technologies into their health care systems. However, often HTAs are constrained with regards to the data upon which these assessments can be made.

HTAs normally have access to data on the efficacy of health care interventions through, for example, Randomised Controlled Trials (RCTs). It is often the case, given that they are built upon manufacture submissions, that only one or two RCTs are used or referenced. As RCTs tend to provide short-term evidence over a controlled population (generally a healthier and younger population than the general at-risk population given trial entry criteria that typically try to minimise the influence of co- morbidities) this data must be converted into life-time health benefit through extrapolation. Similarly, lifetime treatment costs are unavailable and must be extrapolated. The recent emergence of immune therapies, with their potentially curative effects, introduces additional difficulties in making these extrapolations.

In this report for WP2, we concentrate on the methodology applied to estimate the impact on survival estimates used within the incremental cost-effectiveness ratio calculations that underpin many of the value for money decisions used within HTAs. In particular, we focus on estimation and extrapolation of survival benefits although we also give a preliminary evaluation of quality of life methods. We evaluate cost methodologies, which are reviewed in a future, accompanying report.

Extrapolation can be performed using a number of different techniques and even within one of the more commonly used techniques (parametric modelling) there are a number of alternatives.

Alternative approaches to extrapolation can lead to different estimates of survival. Which can impact decisions on health technologies. Even small changes in estimates of survival can be consequential. For example, the estimates of survival are used in the denominator of the incremental cost effectiveness ratio (ICER) – a ratio used to guide technology decisions – and as such a small change in the estimate can give rise to big differences in the ratio, and thus have significant implications for the decisions. For example, in one assessment the use of a Gompertz distribution rather than a Weibull increased the ICER from £13,000 to £23,000 for an assessment of rituximab for leukaemia (1). NICE typically recommends treatments that have ICERs below £20,000 to £30,000, consequently this change could have affected the decision, particularly as NICE takes account of the impact of uncertainty on ICER estimates. As the choice of model can affect the estimates on which decisions are made, failure to systematically chose a model can create the possibility of bias and inconsistency between HTAs and thus inconsistent and inefficient decisions.

Several frameworks - which are discussed in the results - have been presented to justify model choice. This includes Latimer 2013 (2), which was produced by the Decision Support Unit of the National Institute for Clinical Excellence (NICE). Latimer 2013 reviewed HTAs submitted to NICE between 2000 and 2009 and found that none of the assessments included full and systematic justifications for the chosen parametric models, in some HTAs multiple justifications were used, but in all cases flaws remained[4]. Six years after Latimer 2013, it is an

opportune time to consider if the justification of model choice has improved.

This report attempts to improve understanding of the current approach being used to select models for extrapolation, and to critique existing frameworks. The focus is on extrapolation as rarely is clinical evidence full enough to determine final clinical outcome, and therefore extrapolation has to be used to estimate the full population impact of a new health care technology. This report, it is hoped will help improve consistency in model choice, improving HTA and outcomes.

The review feeds into the development of the WP DICE platform, which aims to develop an operational computer simulation model for HTA use in assessing ICERs. We concentrate therefore on the usefulness and applicability of extrapolation models within this context. Specifically, the focus of this report is on the extrapolation of survival benefits. In particular, we are interested in how extrapolation feeds into Markov cohort models, which are commonly used in HTA submissions.

Although it should be noted that the DICE platform which is also being developed as part of WP1 provides a wider and richer set of structural model choice than merely this Markov cohort approach.

To improve the understanding of the current approach to model selection for benefits we first describe and compare the most commonly used methods for the extrapolation of benefits; then provide a brief introduction into commonly used techniques for guiding model choice; then provide an overview of current literature which provides guidelines or frameworks or assesses model choice; and finally compare the methods used, and justification provided in a subset of HTAs submitted to NICE. We also give a concise overview of the manner in which quality of life (QoL) adjustments are made to estimate and incorporate Quality Adjusted Life Years (QALYs) into Markov cohort models, a prevalent form of HTA output and one promoted by NICE amongst other HTAs. From this a discussion on the justification of model choice is presented which is compatible with the development of the DICE platform, which is one of the primary aims of undertaking this review – namely to support the use of the DICE platform by HTAs.

2. Background – Commonly Used Methods For Extrapolation

There are many techniques which can be used to extrapolate survival benefits for HTA. These methods can be classified as parametric models, non-parametric models, or semi-parametric models. Parametric models are the most common; a model is said to be parametric if an assumption regarding the distribution of the survival function (the probability of an event beyond a given time, say t) is made. If no distribution is assumed the model is non-parametric, and if there is a combination of assumptions, the model is semi-parametric.

Possibly the most common non-parametric approach is to consider RCT survival curves and their underlying hazard rates, and then to choose a hazard rate, based on the within trial period to extrapolate forward the cumulative survival curve. The cumulative survival is driven by the underlying hazard rate, which may be crudely thought of as a given periods mortality rate, and this is estimated in non-parametric approaches through counts of deaths at any period in time divided by the total population at risk adjusted for censoring. For example, the

investigators may choose the average hazard rate observed across the RCT period, the last period hazard rate or some other hazard rate assumed to hold going into the future to extrapolate results over lifetime. Alternatively, investigators may assume the difference in observed hazard rates within the RCT period continues through lifetime until a pre-specified end period. Or the observed hazard rates are matched to existing observational hazard rates gained from pre-existing observational or life-table data. This later method was used in early models such as the extrapolation of the GUSTO trial (3) data or the WOSCOPS data (4) for economic analysis. The problem with all these non-parametric approaches is that they are entirely reliant on investigator assumption of best fit, do not lend themselves to statistical inference and are consequently highly subjective. For this reason, most HTA submissions have moved to more sophisticated parametric techniques. As such, we do not consider non-parametric estimators further.

Semi-parametric techniques are commonly associated with the fitting of proportional hazards (commonly referred to as Cox proportional hazards) models that assume that the difference between the intervention and control study population hazards of mortality are proportionally different throughout the study period, and if extrapolated continue to be so. Obviously, this is a relatively strong assumption reflecting a form of treatment benefit that may hold for some period of time but is unlikely to be continuous. While these semi-parametric models can incorporate a number of risk factors (useful for sub-group analysis in particular) into their specification and while still commonly used within trial settings are rarely used to extrapolate survival forward given that hazard proportionality is a strong assumption to hold over lifetime.

That said, recent developments in the analysis of within trial data have emphasised the fit of models to observed data, whether this be proportional or non-proportional, through the fitting of parsimonious non-parametric models to the observed data. This has followed the seminal work of Royston and Parmar (5) in providing estimators of a wide range of piecewise parametric models, based on fitting the observed data through imposed linear splines, that also encompass (nest) various semi-parametric and parametric functional forms within their general specification. These Royston-Parmar models also embed extrapolation functions, as this is relatively simply built of the linear spline approach, into their approach. However, as the extrapolation essentially is built off the last linear spline fitted to the data it entails an assumption that this spline of the hazard progresses into the future. Even though the linear splines can encompass a number of assumptions regarding the underlying hazard, which are made even more generalisable through altering the modelling of time within their models, in terms of extrapolation this is as strong an assumption as that encompassed by the parametric approaches. Thus, while the Royston-Parmar approach has generalised the analysis of within trial data and makes the extrapolation into future periods relatively straightforward, it must be used with some caution.

As such, we concentrate on the more commonly used parametric models that have gained attraction in extrapolating survival benefits. Whether non-parametric, semi-parametric or fully parametric all models have the ability to calculate the average survival gain associated with a treatment intervention. This overall average treatment gain is the estimated mean survival time gained from the estimated extrapolation. This is normally calculated as the expected survival time from a parametric or semi-parametric model and through estimation of the "area under the curve" in non- parametric models. As the modelled extrapolation changes, so do

the estimates of average survival time change. This is sometimes referred to in the literature as model uncertainty. It is important to document the impact that such changes can have on expected survival time, especially as these estimates form the basis of the denominator in the ICER.

Parametric models are the most used type of model in HTA because they are useful when right-hand censoring exists. Right-hand censoring occurs when an event of interest did not occur during the observed period, and thus the time to event is unknown. This is common in RCTs, as the prespecified endpoint may not be attained by the trial end. Whilst non-parametric distributions may be able to fit the underlying data from the RCT better, in terms of extrapolating the data if a distributional assumption is defensible, parametric approaches will give more reliable extrapolated estimates.

Parametric distributions are also relatively convenient, in that equations that translate the model parameters into transition probabilities are well-defined. This simplifies the modelling process. Although as will be noted, justification for the adopted parametric distribution is not always provided.

There are a wide range of parametric model forms, with each having their own characteristics. This can make them suitable for different data sets, but also implies that suitable criteria for choice must be given. The more commonly used distributions in HTA are described in Table 1 and in Section

2.1.1. There can be similarities between these distributions, with several distributions nesting other distributions. For example, the Weibull, Log-logistic, and Gamma distributions are compatible, under specific circumstances, with exponential distributions.

Where there is no censoring, a rare occurrence in HTAs, the use of non-parametric models is more appropriate. Non-parametric models make no assumption regarding the underlying distribution of the time to event, are fully defined by the empirical distribution. There is a limited need to make an assumption when there is no censoring within the data. The most used non-parametric model in HTA is the Kaplan-Meier model.

Given that censoring is commonly encountered, the most common types of models used within HTA submissions, again given the reliance on RCT data, are proportional hazards models and accelerated failure time models. In a proportional hazards model the hazard rate for the technology is a proportion of the hazard rate for the control. Generally speaking proportional hazards models are defined as semi-parametric specifications. In an accelerated failure time (AFT) model the hazard rate for the technology may be defined to increase or decrease over time. Approaches using AFTs are normally specified as parametric models.

2.1 Parametric model forms

When parametric models are used the distribution chosen can either be applied to the entire dataset with the treatment group included as a covariate, or the dataset can be split and individual parametric models can be applied to different time periods. If the later approach is taken this is a piecewise parametric model. This approach provides a simple technique to model a variable hazard function and is more flexible than individual parametric models. There are also more general weakly structured, flexible piecewise parametric models available

– such as the Royston and Parmar spline- based models as noted above (5, 9).

Another challenge with survival analysis is the influence cofounders can have on survival. Cofounders are factors that can affect survival, for example gender. If they are not accurately accounted for, they will bias the extrapolation. Cofounders can be included in survival analysis, through for example, Cox-proportional hazard models. The Cox proportional hazard model is a multivariate statistical model which allows the effect of these factors to be evaluated as well as accounted for as noted in passing above.

2.1.1 Commonly used distributions in parametric modelling

The most simplistic distribution is the exponential. The defining feature of this distribution is a constant hazard function and as such it can be defined by one parameter. This also means that it doesn't have memory giving it a 'no-ageing' property. This property reduces its possible applications as few 'time-to-events' have this property. It has been used in the simple Declining Exponential Approximation of Life Expectancy (DEALE) method (6). This is a method used to approximate the life expectancy for an individual patient using information from various sources, such as disease-specific survival rates and age- and sex-specific life expectancies from a table of vital statistics. The exponential distribution is essentially a proportional hazards model. As noted above this assumption in itself limits its use for extrapolation purposes. That said, many software programmes rely on exponential decline in the survival rate to model small numbers or censoring time at the end of study period to ensure the estimated model returns non-censoring at the final study time period, that is, that all within a parametric modelling process all individuals remaining at the study end reach the endpoint.

Unlike the exponential, the Weibull distribution can have an increasing, decreasing or constant hazard function dependent on the parameter values chosen to specify the model. The function is still monotonic, and as such it cannot change direction within the study timeframe. It is worth noting that a Weibull distribution with a constant hazard function returns an exponential distribution.

Weibull distributions can be parameterised either as a proportional hazards model or as an accelerated failure time model. This is an important feature. It means that the technology can act proportionally or accelerate/decelerate over time compared to the control. The Weibull also has a shape and scale parameter, allowing it to be quite flexible. The shape parameter affects the slope of the line in a probability plot (the probability density function), whilst the scale parameter stretches the distribution along the (in survival analysis) time axis. The Gamma distribution is similar to the Weibull, but not as mathematically manageable. The distribution includes the exponential as a special case (when the scale parameter equals one). The Gamma can be used as an AFT model.

The Generalised Gamma distribution is a flexible three-parameter model, and is a generalisation of the Gamma distribution. It is useful because it includes the Weibull, exponential and log normal distributions as special cases. As such it can help to identify whether a Weibull, Gamma or log normal model may be suitable for the observed data. The Generalised Gamma can be used for AFT models.

The Log-logistic distribution is similar to the Weibull and exponential models in that it specifies the simple expressions for the hazard function and the survival function. This distribution can have a non-monotonic hazard function, for example an initially increasing hazard, followed by a decreasing hazard. Consequently, when considering the applicability of the log-logistic distribution the validity of non-monotonic hazards must be considered. This characteristic is also why it is the most commonly used distribution for AFT models. It cannot be used for proportional hazards models.

The Gompertz distribution has two a shape parameter and a scale parameter. It increases or decreases monotonically. The hazard can have an exponential distribution, can increase or decrease monotonically with time. It does however differ from the Weibull because it has a log-hazard function which is linear with respect to time, whereas the Weibull distribution is linear with respect to the log of time. The Gompertz distribution is only applicable if monotonic hazards are appropriate. It is suitable for proportional hazard models.

The normal and the log-normal distributions are specified by two parameters (mean and variance) and follow the normal or log-normal distribution. The hazard rate of the log normal is hump shaped. It starts at zero, increases, returns to zero, subsequently the hazard function is declining for large values of X. This form of the hazard function is generally considered to be implausible in HTA, and as such it is not widely used. The log normal distribution is very similar to the log-logistic distribution. The similarities between the logistic and normal distributions mean that the results of log-logistic models and log normal models are likely to be similar. As with log-logistic models, when considering the applicability of the log normal distribution the validity of nonmonotonic hazards must be considered, and the validity of potentially long tails in the survivor function must be considered. The log-normal distribution can be used in AFT models, and not for proportional hazards models.

Other distributions, such as the Exponential power, Inverse Gaussian, and Pareto exist but are less commonly used than the aforementioned distributions. This is largely as they have been found empirically to be less useful than the Weibull and Gompertz distributions in particular when fitting to specific study data. Therefore, they are not discussed further.

The hazard rate function and the mean estimates, noting that the latter provide estimates of mean survival time returned from these parametric models, of these various distributions are provided in Table 1. As is easily seen from the Table the parameterisation of these distributions, and the resulting survival and hazard rates are based on markedly different estimates. Given that it is the hazard rate that defines the underlying approach to the extrapolation this is of crucial significance. Although as noted, by placing restrictions on parameters in certain cases, a relationship may be established between different functional forms.

Table 1. Hazard Rates, Survival Functions and Expected Value for Common Distributions

	Hazard Rate h(x)	Survival Function S(x)	Mean E(x)
Exponential λ > 0; x ≥ 0	λ	$e^{-\lambda x}$	$\frac{1}{\lambda}$
Weibull α, λ > 0; x ≥ 0	$\alpha\lambda x^{\alpha-1}$	$e^{-\lambda x^{\alpha}}$	$\frac{\Gamma(1+\frac{1}{\alpha})}{\lambda^{1/\alpha}}$

Gamma β, λ > 0; x ≥ 0	$\frac{f(x)}{S(x)}$	1 -*I*(λx, β)	$\frac{\beta}{\lambda}$
Generalised Gamma λ, α, β > 0; x ≥ 0	$\frac{f(x)}{S(x)}$	1- *I*[λex^{α},β]	$\int_0^x S(x)dx$
Log-logistic* α, λ > 0; x ≥ 0	$\frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$	$\frac{1}{1+\lambda x^{\alpha}}$	$\frac{\pi Csc(\frac{\pi}{\alpha})}{\alpha\lambda^{1/\alpha}}$
Gompterz θ, α > 0; x ≥ 0	$\theta e^{\alpha x}$	$e^{[\frac{\theta}{\alpha}(1-e^{\alpha x})]}$	$\int_0^x S(x)dx$
Normal σ > 0, -∞ < x < ∞	$\frac{f(x)}{S(x)}$	$1-\phi\frac{x-\mu}{\sigma}$	μ
Log-normal σ > 0; x ≥ 0	$\frac{f(x)}{S(x)}$	$1-\phi\frac{1nx-\mu}{\sigma}$	$e^{(\mu+0.5\sigma^2)}$

f(x) is the Probability Density Function

*If α > 1

2.2 Criteria to justify model choice

As there are a variety of models available for extrapolation it is necessary to choose and justify the choice. There are several criteria that could be considered when justifying model choice. Some relate to internal validity, how well the model fits the data available, while some relate to external validity and how plausible the model is for extrapolated data.

Criteria to assess internal validity include goodness-of-fit and visual inspection. Whilst clinical validity; and external data can be used to assess external validity. Visual inspection requires a plot of the observed hazard rates over time. For examples plots of log cumulative hazards against time or against log time. These plots can help us to understand whether hazards are likely to be non-monotonic, monotonic or constant. This property will guide the choice of distribution or model. The plot can also be used to identify if the proportional hazards assumption holds. However, there are challenges with using a visual inspection. It only reflects data from the trial and personal perspectives or judgment is required. This can mean that modellers arrive at different conclusions when considering the same data. Personal judgment is removed when statistics measuring goodness-of-fit are used.

There are several alternative statistics to measure goodness-of-fit. However, the most commonly used statistical measures of goodness-of-fit are the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These tests weigh up the improved fit of models, with the use of additional parameters. The BIC penalises the use of additional parameters more strongly than the AIC. There are other less commonly used statistical tests available, for example the Cox-Snell residuals (7), which assess how closely a parametric function follows the Kaplan-Meier function.

Additional options include splitting the observed data at random, developing a model based upon one portion and evaluating it on another, and k-fold cross validation and bootstrap resampling (8).

A challenge with all statistical measures of fit, is that they only consider the data from the

trial. This may not be a concern if there is relatively little censoring and if the data are generally complete. However, if this is not the case, the plausibility of the fit can be important. This can be assessed through the use of external data or expert opinion.

If external data are available from an alternative clinical trial or from a long-term registry data it could be used to assist with model choice and justification. External data can be useful in assessing the plausibility of the extrapolation. Whilst it is preferable that the external data is patient level, aggregated data can also be useful as it can, for example, indicate what an appropriate overall survival rate after a period of time could be. A challenge with using external data for model choice it that external data is likely to only be available for the control, due to the fact that the technology being assessed is normally novel and new. As such it is likely that external data is helpful for the extrapolation of the control, but less helpful for the technology. Furthermore, even if external data is available, efforts should be made to ensure the external patient level data reflects patients similar to those in the HTA. There are methods are available (10) to validate a fitted model using an external dataset.

Expert opinion and clinical plausibility can also be used to guide model choice. This is crucial to ensure the model is appropriately specified and extrapolated results are feasible. Models chosen for extrapolation should provide results that are clinically valid, justifiable and biological plausible.

3. Methods

We consider the justification of model choice from a theoretical and practical perspective. The theoretical perspective includes a review and critique of notable published frameworks and guidelines on model justification. The practical perspective considers the extent to which models are justified and how they are actually justified in practice.

To understand how systematic the approach is to model choice, we focus on and compare model choice in HTAs submitted to NICE for metastatic cancer between 2015-16 and 2019. We also use NICE appraisals as they are publicly available, reported in depth and have a clear methodological guidance associated with the acceptable modelling decision. The time frame of 2015-2019 was chosen as the data sources (the company HTA submissions to NICE) were consistent over this time period, and this reflected a unique sample for assessment. To maximise comparability whilst maintaining a feasible sample of HTAs, we focused on HTAs of treatments for metastatic cancer and on the use of extrapolation within Markov cohort models (a frequently used model in HTA where future states depend only on the current state). We also use NICE appraisals and the Markov model as these are of practical importance to our validation of the DICE platform, which forms the central component of the WP.

Metastatic cancer is chosen partly because of the recent debate (8) over the use of overall survival versus progression-free survival as an end-point for both trials in this area and extrapolated models. Progression-free survival (PFS) has become a frequent outcome used to evaluate new cancer drug efficacy. However, some research has failed to find a substantial link between PFS and health-related quality of life (HRQoL) in cancer trials, casting doubt on its role as a surrogate endpoint.

Furthermore, PFS benefits do not always translate into overall survival benefits. HTAs submitted to NICE were chosen as NICE is one of the earliest agencies to focus on the use of cost-effectiveness in assessing new health care technologies and has developed a clear methodological stance in the assessment of evidence. The NICE Decision Support Unit provides technical support documents to assist with HTA. This includes the study by Latimer in 2013 'Survival analysis for economic evaluations alongside clinical trials – extrapolation with patient-level data' (2).

HTAs which are updates of previous ones, and HTAs which have been terminated have been excluded from the review. To ensure we understand how companies justify their model choice we also only consider the first HTA submission for each technology. This excludes any additional information, modelling or analysis as requested by NICE, and thus ensures our review reflects company justifications, rather than NICE's.

4. Results

4.1 Published frameworks to justify model choice

There are several published frameworks or algorithms to guide model choice and help ensure decisions are made systematically. Arguable, the most notable of these is Latimer 2013. The provision of guidance is of high methodological importance - due to HTAs facing increasing issues concerning the use of real-world data and its incorporation into CEA and guideline assessments. A summary of selected literature discussing frameworks and guidelines is below.

Latimer 2013 (2)

In a review of 45 HTAs submitted to NICE between 2000 and 2009, the survival models chosen were not systematically justified in any of the assessments. Rationale for the chosen model was only provided in a limited number of HTAs and very rarely was consideration given to the plausibility of fit for the chosen model.

Latimer proposes a systematic approach for choosing survival models (a model selection process algorithm) and concludes that justification should be based on the fit of alternative models to the observed data and the clinical plausibility of the extrapolated portion of the curve. If censoring is not present the Kaplan-Meier approach is suggested. If censoring is present log-cumulative hazard versus log of time plots are suggested as a means of validating within trial survival and AIC and BIC tests to differentiate fit. Latimer proposes that all ''standard'' parametric models (exponential, Weibull, Gompertz, log-normal, log-logistic) should be considered and demonstrated to be suitable or not. If not, more flexible models may be required. Furthermore, alternative plausible models should be considered in sensitivity analysis.

Latimer also notes that external data can be a useful criterion to understand external validity, and whilst there are limitations (e.g., compatibility with the trial in question and lack of evidence) an attempt at justification is better than none.

Connock 2011 (11)

Using examples of HTA from drugs used to treat cancer Connock concluded that function fitting to observed data should not be a mechanical process validated by a single crude indicator, i.e. the AIC. Projective models should show clear plausibility for the patients

concerned and should be consistent with other published information. Multiple rather than single parametric functions should be explored and tested with diagnostic plots. When trials have survival curves with long tails exhibiting few events then the robustness of extrapolations using information in such tails should be tested.

Furthermore, Connock found that, the choice of model had a large effect on the predicted treatment-dependent survival gain, for e.g. logarithmic models (log-Normal and log-logistic) delivered double the survival advantage that was derived from Weibull models.

Jackson 2010 (12)

Jackson compared the fit and predictive ability of parametric and semiparametric models by using the deviance information criterion to account for model uncertainty in the cost-effectiveness analysis. Under the Bayesian semiparametric models, some smoothing of the hazard function is required to obtain adequate predictive ability and avoid sensitivity to the choice of prior. Jackson found that one flexible parametric survival model fits substantially better than the others considered.

Diaby 2013 (13)

Diaby proposes a tutorial for the application of appropriate survival modelling techniques to estimate transition probabilities. Specifically, researchers can reconstruct the patient data from published KM curves, using an algorithm implemented in the statistical package R. In selecting the best parametric model to fit their data, researchers should use both statistical and graphical tests. Parametric survival modelling techniques are suitable for developing equations for transition probabilities for use in model-based economic evaluations.

Guyot 2011 (14)

Guyot reviewed the survival analysis undertaken alongside 24 CEAs in HTAs undertaken for NICE between June 2008 and September 2014. They found that most often, proportional hazard modelling was undertaken. However, noted two important problems with the analysis. None of the studies tested the proportional hazards assumption and the source of the hazard rate used within these analyses was not specified—which may cause bias if it was derived from an unrelated model.

Bagust and Beale 2013 (15)

This study derives directly from Latimer (2) and highlights that the Latimer suggestions are premised on the availability of individual patient data to allow assessment of the survival analysis and extrapolation within an HTA setting. While this data is normally available to an HTA, the authors point out it is rarely publicly available, for reasons of commercial confidentiality. While this may be the case it might be argued that any published data should be made publicly available. They suggest observation of the cumulative hazard plot to outline overall trend in the data. They also suggest close examination of the clinical protocol to highlight any differences affecting the composition of the trial population and the more general treatment population. They also suggest explicit sub-group analysis, particularly where the trial shows high early mortality with some prolonged survival benefits in sub-groups of individuals, as is increasingly commonly seen in trials of patients with malignant

melanoma. In providing some detailed principles, they suggest that extrapolation must be supported by association with likely modes of treatment benefit.

Gorrod et al 2019 (16)

This review of NICE HTA studies in the area of oncology covers similar ground to our own review, reviewing studies between 2011 and 2017. Looking at 58 studies they find the majority use parametric models and statistical assessment as the basis of model choice. They highlight the need for clinical assessment and display of the underlying within trial hazard to justify any specific method chosen for extrapolation, and that model selection was not undertaken in a systematic manner.

4.2 Review of NICE HTAs

We identified 47 HTAs reviewed by NICE between 2015-16 and 2019, which used Markov cohort models to assess technologies to treat metastatic cancer.

4.2.1 Distributions chosen in HTA submissions

There is significant variability in the models chosen for survival analysis. Around a third (16 of the 47) of the HTAs used the exponential distribution, nearly a quarter used a Weibull distribution, around a fifth used the log-logistic, around 10% used the log-normal, and a further 10% used the generalised gamma (table 2). Only two HTAs used non-parametric models.

Table 2. Distributions used in HTA submissions (2015-16 to 2019)

	Exponential	Weibull	Log-logistic	log normal	Generalised Gamma	Gamma	Non-parametric
Count	16	11	9	4	4	1	2
Percentage	34%	23%	19%	9%	9%	2%	4%

It is not surprising that a range of models were used in the HTA. Different models will be appropriate in different circumstances and contexts. For example, different distributions may be appropriate for different primary outcomes. As the primary outcome differed (either overall survival or progression free survival was generally used) it may be appropriate for the model to also differ. Concerns around the use of different models arise when there is inconsistency in choosing the distribution. To consider this we reviewed the model justification provided in the HTAs.

4.2.2 Justification of model choice

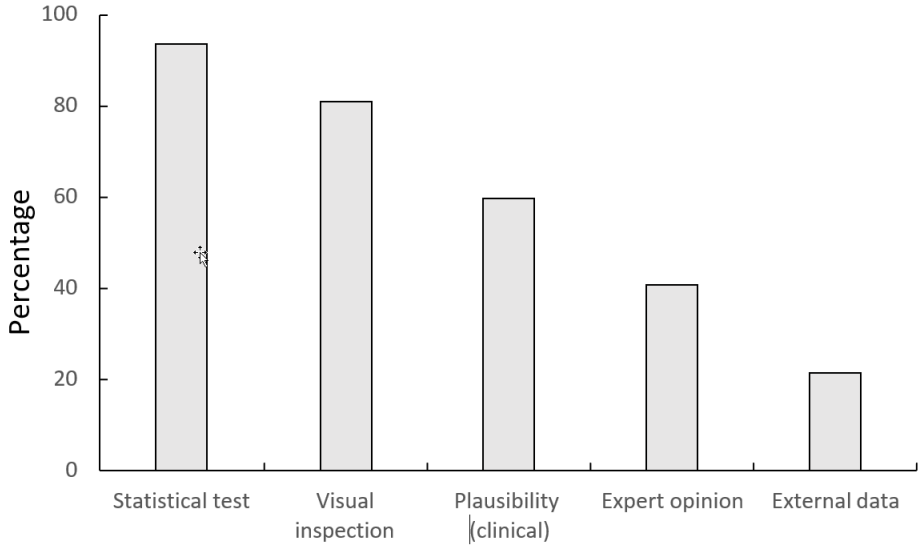
All but two of the 47 HTAs reviewed provided some details of the justification process. Statistical tests were used in all but three HTAs (94%) and visual inspections were used in 38 of the 47 HTAs (80%). Of the three HTAs that did not use statistical tests, this included the

two that did not provide any justification for the extrapolation and one HTA where visual inspection was used to determine that parametric models were not appropriate, and as such statistical tests were not used.

The plausibility of the distribution is a factor in model choice in 28 of the 47 (60%) of HTAs, and expert opinion was used in 19 of 37 (40%) HTAs (Chart 1). It is, however, quite feasible that experts provided opinion on the plausibility of the extrapolation model yet this was not explicitly declared or discussed in the HTA submission. It is also not clear how much benefit there is to consider clinical plausibility if it is provided by a non-expert in that specific clinical area. There were also three submissions where an expert was consulted, but the plausibly of the extrapolation was not explicitly mentioned as a consideration in their deliberation. This appears to be an odd use of the expert.

The use of external data is relatively limited, being used in only ten (21%) of HTAs. The submissions that used external data and the external data used is presented in Table 3. It's unclear what drives the low usage of external data. It could be due to data availability, technical expertise, or even awareness of the technique. What is clear is that many HTAs therefore have low external validity.

Chart 1. Use of model justification techniques in HTAs (2015-16 to 2019)



5. Discussion

5.1 Review of HTAs submitted to NICE

The overview of model justification in HTAs reviewed by NICE between 2015-16 and 2019 provided several insights. Consistent with findings of Latimer 2013 (2), a variety of models are being used to extrapolate survival benefits. This does not suggest incorrect modelling is taking place. Rather this may be appropriate as different models may be appropriate in different contexts.

Justification of model choice has improved significantly since Latimer 2013 (which assessed HTAs between 2000 to 2009), but remains non-universal. The increasing use might be due to the framework provided in that paper, Latimer 2013 (2) was referenced in at least a dozen of the HTAs reviewed. However, despite a significant increase in model-choice justification, there was still disagreement between the NICE Estimates Review Group (ERG) and company submissions for a number of models chosen. The proposed technology was not recommended in at least six of the 47 technologies we reviewed, although this was not always due to model choice however.

Our view from interpreting the recommendations made by NICE is that clinical plausibility was one of the main reasons for the proposed extrapolation model to be rejected. Clinical plausibility is integral to model acceptance and a more binary criteria than visual inspection or statistical tests.

Visual inspection was also a source of disagreement between the ERG and the company. As visual inspection requires judgment different modellers can have different perspectives of the best fitting model. Thus, whilst visual inspection can assist with model choice, it does not always provide a systematic approach to model choice. It is this systematic approach which is crucial to ensure consistency between HTAs. In many instances, the final position of NICE was that there is considerable uncertainty and a range of possible models should be considered.

There were also instances where the NICE ERG disagreed with the views presented by experts. This is another challenge which arises due to the use of judgement, and judgement in relatively complex and niche topic areas.

5.2 Review of model choice frameworks

All the reviewed frameworks mentioned that the choice of model is important and that a range of parametric models should be considered. They all include statistical tests and visual inspection as considerations for model choice.

Amongst the frameworks reviews there is limited discussion regarding the relative importance of each of these attributes. This presents a challenge as a best fitting model under one criteria may be a poor fitting model under another criteria.

Latimer 2013 was referenced in a number of HTAs reviewed and the model selection algorithm appears to have been followed. As such we consider that it has been influential.

One of the challenges to consider is the trade-off between internal and external validity and prioritising between the different attributes used for justification. Model choice should be guided by both internal and external validity, but differing weights given to these two criteria will affect model choice. Clearly given the availability of any long-term external data for the control comparator there is a clear, basic and robust criteria that could be applied to any chosen model; does it extrapolate the control group in line with long-term external data?

At a more granular level, how do we prioritise the model fit attributes. For example, is a low AIC more important than visual fit? Furthermore, decisions that involve judgment can also

complicate the prioritisation of the different attributes. There are even challenges with the less subjective tests. Ranking of AIC and BIC statistics doesn't provide a perspective on relative difference. A model may be the worst, but this doesn't provide a perspective on how much worse. A framework or choice architecture could be beneficial toward answering this type of question. An alternative would be to include a step-wise rejection process, whereby models are rejected if they are not suitable given a hierarchy of criteria. The full suite of feasible models could then be considered in the HTA, or at least included in a sensitivity analysis.

5.3 Limitations

There are many methods available for survival analysis. As such, this report is not a comprehensive examination of the estimation and extrapolation of survival analysis as applied to health benefits. It is also not intended to be. Rather we are simply attempting to describe the most commonly used methods in HTAs, with a view to discussing the use of criteria to choose between them, with a view to aiding the development of our WP platform to undertake HTAs modelling.

Our review of model justification in HTAs is reliant on information provided in NICE company submissions. If there were factors that affected model choice which were not sufficiently (or explicitly) detailed in the HTA submission, they would be inadvertently excluded from this review.

5.4 Issues arising for extrapolation associated with immuno-therapies

The new genetic based technology for finding treatments has had a far-ranging impact on R&D and on the establishment of efficacy. Rather than identifying a target mechanism for a disease and designing a drug around that mechanism, immuno-oncology based on genomics has generally adopted an inductive approach. In developing immuno-oncology treatments, the cellular responses produced by specific agents are analysed using statistical algorithms. The approach uses principal- component type models to examine correlations between large numbers of potentially predictive biomarkers and cellular responses. This involves, for example, examining SNPs across many proteins at one level and many molecular structures at another to determine which combinations of proteins and molecules are affected by the specific monoclonal antibody (MAB) to produce an immuno- response that attacks cancer cells. It is, therefore, based on an identification of statistical correlations which may identify causal pathways. Of course, the correlations may not be reflective of causal effects.

If a significant immuno-response effect is found, the particular combination is then tested in a small number of patients. If the effect of the MAB is maintained in patients, then testing enters the R&D Phase II and, if results at this point are positive, the rest of the testing process may proceed under accelerated approval procedures. This shortens the review process and therefore the time required to gain marketing approval.

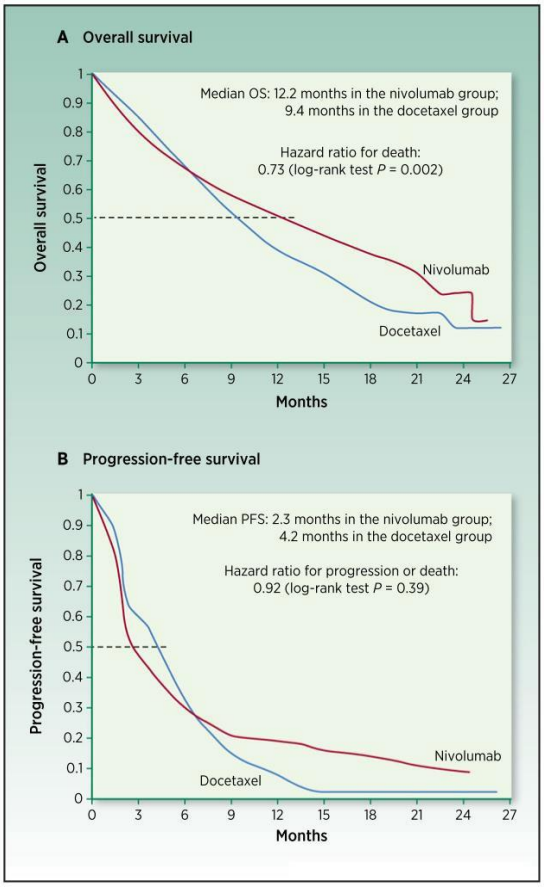
Clinical research is not quite the same for immuno-oncology agents as for traditional cancer therapies and a positive effect is somewhat open to interpretation. Responses to immuno-oncology agents are generally based on their effect on immune response, rather than the direct killing of cancer cells. Determining the extent of effect takes longer because the immuno-response takes some time, which means that either clinical trials must be longer or

intermediate endpoints must be used. The process is further complicated because it can be difficult to assess response using standard measures of positive response—e.g. of tumour size, and patients respond differently, as noted above—some individuals may be responders, while some remain non-responders.

Tracking effectiveness over time also can be a challenge. Overall response rates are commonly defined as early indicators of treatment success, but response may not be long lasting. Toxicities may arise due to autoimmune activation which are unpredictable in terms of timing and severity and may trade-off any observed short-term response rates. progression-free survival may last for a time, then the cancer reappears. Surrogate markers, moreover, may not correlate well enough with actual clinical outcome. The complexity and uncertainty of monitoring and demonstrating positive effect is evident in the variation among regulatory agencies in accepted endpoints. Disagreement continues over what constitutes adequate tumour response and how to measure different levels of response. Another layer of complication is added when the same therapy is licenced for different indications with various outcomes.

Data analysis in clinical trials is also more complicated for MABs. In traditional clinical trials, data usually are analysed using proportional hazard models, which focus on treatment differences across control and intervention groups. An important purpose is to better define response in subpopulations. However, for some gene-based therapies, response is not necessarily proportional and a small number of responders may have long-lasting effects. A non-proportional response may mean that the cure effect increases over time, an outcome that can be know with certainty only if all patients are followed for a long time.

Figure 1 Survival analysis in immune-oncology example (Karzandjian et al, 17)



To use an actual therapy as an example, Kazandjian et al. (17) compares overall survival for nivolumab and docetaxel for non-small-cell lung cancer (see Figure 1). Note that early on nivolumab is not as effective as docetaxel because it takes time for the immuno-response to have an effect, a common finding in this area. Graphs for progression-free survival for these treatments show a similar cross-over. What is clear is that response is not proportional, so traditional statistical analyses (based on proportional hazards) may not be appropriate and data must be collected out far enough to capture any true positive treatment effect.

Other analytical challenges are presented by this general statistical approach. For example, the appropriate "p-value" used to differentiate the effect in the treatment and control populations will be affected by the number of hypotheses that need to be tested in using SNP algorithms to identify response correlations. In clinical trials, proportional hazard tests usually employ a logrank test, similar to a chi-squared test versus expected values, to test for differences in survival curves.

Multiple testing requires adjustment to such tests. Statistical approaches also have to account for the lag phenomenon. Developments in this area are relatively recent. One common adjustment is through the Karrison test (18)**,** accepted as valid only in 2015, which seeks to display true positives when an effect is delayed. Trials also are changing as more therapies are being developed as combinations, which presents additional challenges in recommending dosage, timing and frequency of treatment.

All such matters are currently under consideration, at least academically. They will have a profound impact on HTAs. It is clear to see that extrapolation of effects becomes even more complex when dealing with immuno-therapy as treatment success may be reliant on a sub-population if individual responders, given that treatment is partially a function of autoimmune response, who may only be identified after a considerable time as the autoimmune response is normally subject to delay. This can cause the treatment to perform badly relative to a control in early periods but become increasingly effective, in an uncertain number of individuals, in late phase. Clearly justification of extrapolation of survival methods under such circumstances, although increasingly common in the area of oncology, becomes increasingly complex.

6. Quality of life concerns

Although inherently complex health outcomes within Markov cohort models the multidimensional aspects of health outcomes are compressed into a number of disease states through which individuals move. Generally, they move with no individual history – Markov models are memoryless. Typically, quality of life weights are applied in Markov cohort models to given states. Markov models based on disease states are normally evaluated in discrete time determined by the cycle length, that is the time it is specified by the model for individuals to move between disease states. The use of discrete time assumes that changes in health states, and therefore quality of life weights, occur only at the end of a cycle period. These discrete-time Markov models can only ever approximate the process of disease progression and quality of life changes, as clinical events typically occur in continuous time.

For example, as individuals move through the various disease states specified by the model a quality of life weight is applied to the state, with the same weight applied to the duration of individuals within any given state. This is obviously a crude manner of extrapolating quality of life but is relatively common in application give both the sparsity of data for given diseased states and the ease through which such weights can be applied in this manner. Given the assumption of a lack of memory that characterises Markov cohort modelling, this means that the quality of life weights are not dependent on the prior disease pathway. An individual entering a disease state of late-stage cancer from prior perfect health is given the same quality of life weight as an individual entering the state of late stage cancer from a prior state of early stage cancer for example. Little work has been undertaken on the empirical validity of this modelling assumption (19)**.** Neither has much work been undertaken of the impact of state duration on quality of life weights. Duration issues are a particular concern where the Markov cycle length is long. Work on statistical methods combining quality of life and survival analysis remains in its infancy (see for example (20, 21)**)** and we know of no HTA which has incorporated such methods. Recently, a number of studies have begun to experiment with incorporating quality of life estimates into discrete event models, for which DICE also has functional capability (e.g. see 22) Obviously, both are critical areas for future research and of significant importance in extrapolating quality of life weights over lifetime.

7. Conclusion

The number of HTAs which have justified the choice of model for extrapolation has improved significantly since the early 2000's. This review of HTAs reviewed by NICE since 2015-16 found nearly all included a justification for the model chosen. The significant increase in model justification may reflect

the release of NICE guidelines (2). The framework presented in Latimer 2013 (2) is relatively comprehensive and should be utilised during HTA.

However, there is still opportunity for improvement. Whilst statistical tests and visual inspections are routinely used, expert opinion and clinical plausibility should be used more frequently where possible. In several cases the lack of consideration for clinical plausibility in company submissions was a reason for disagreement between the NICE and the company submission. The use of external data offers opportunity to improve extrapolation. Certainly the control comparator extrapolation or fit would aid external validity of any new treatment extrapolation. It is unclear what drives the low usage of external data in past HTAs. It could be access to the required data, or awareness/capability of the approach. This is an area for further consideration.

HTA submissions should also improve transparency in detailing the justification behind model choice in the extrapolation of survival benefits in HTA. The justification for choosing a model, and the limitations of chosen models should always be explicitly made. It is not sufficient to say a model was chosen as it was considered to have the best fit, without detailing the criteria which defined best fit.

In extrapolating quality of life weights, and therefore QALYs, the methods remain simplistic and crude. The current widespread adoption of Markov cohort models means that treatment history, disease duration and the discrete modelling of time gives rise to weighting that is bound to incorporate error. Use of discrete event simulation models, such as enabled by the DICE approach, at least will allow more sophisticated modelling as it incorporates disease duration and treatment history in a less abstract manner and continuous time pathways. That said, data limitations exist in some disease areas.

8. Recommendations

Our prescriptions going forward are as follows:

- Always outline the cumulative hazard curves in proposing a particular extrapolation and show how the extrapolation builds off this
- Following Latimer, we accept that different parametric models may be used and that reasons should be given for discarding common models
- Explicit justification should be given for the accepted extrapolation model and the limitations, including data limitations, explicitly outlined
- All extrapolations should include clinical justification and explicit outline of the causal treatment effects
- Particular care should be taken when extrapolation immune-therapies and a wide range of sensitivity analysis should be undertaken in these cases
- More explicit and detailed explanation of the assumptions underlying the quality of life weights used to extrapolate QALYs should always be given

9. References

1. leukaemia-chronic-lymphocytic-first-line-rituximab-evidence-review-group-report2.pdf [Internet]. [cited 2019 Aug 26]. Available from: https://www.nice.org.uk/guidance/ta174/documents/leukaemia-chronic-lymphocytic-first-line-rituximab-evidence-review-group-report2
2. Latimer NR. Survival Analysis for Economic Evaluations Alongside Clinical Trials—Extrapolation with Patient-Level Data: Inconsistencies, Limitations, and a Practical Guide. Med Decis Making. 2013 Aug;33(6):743–54.
3. Mark DB, Hlatky MA, Califf RM. Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction N Engl J Med 1995 May 25;332(21):1418-24
4. Caro J, Klittich W, McGuire A, et al. International economic analysis of primary prevention of cardiovascular disease with pravastatin in WOSCOPS. West of Scotland Coronary Prevention Study. *Eur Heart J*. 1999;20(4):263-268. doi:10.1053/euhj.1999.1193
5. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects [Internet]. Statistics in Medicine. 2002 [cited 2019 Aug 26]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1203
6. Beck JR, Kassirer JP, Pauker SG. A convenient approximation of life expectancy (the 'DEALE'). I. Validation of the method. Am J Med. 1982 Dec;73(6):883–8.
7. Cox DR, Snell EJ. A General Definition of Residuals. J R Stat Soc Ser B Methodol. 1968;30(2):248–75.
8. Harrell F. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis [Internet]. New York: Springer-Verlag; 2001 [cited 2019 Aug 26]. (Springer Series in Statistics). Available from: https://www.springer.com/gp/book/9781441929181
9. Royston P, Parmar MKB. External validation and updating of a prognostic survival model. :24.
10. Kovic B, Jin X, Kennedy SA, Hylands M, Pedziwiatr M, Kuriyama A, et al. Evaluating Progression-Free Survival as a Surrogate Outcome for Health-Related Quality of Life in Oncology: A Systematic Review and Quantitative Analysis. JAMA Intern Med. 2018 Dec 1;178(12):1586.
11. Connock M, Hyde C, Moore D. Cautions Regarding the Fitting and Interpretation of Survival Curves: Examples from NICE Single Technology Appraisals of Drugs for Cancer. PharmacoEconomics. 2011 Oct;29(10):827–37.
12. Jackson CH, Sharples LD, Thompson SG. Survival Models in Health Economic Evaluations: Balancing Fit and Parsimony to Improve Prediction. Int J Biostat [Internet]. 2010 Jan 27 [cited 2019 Aug 26];6(1). Available from: https://www.degruyter.com/view/j/ijb.2010.6.1/ijb.2010.6.1.1269/ijb.2010.6.1.1269.xml
13. Diaby V, Adunlin G, Montero AJ. Survival Modeling for the Estimation of Transition Probabilities in Model-Based Economic Evaluations in the Absence of Individual Patient Data: A Tutorial. PharmacoEconomics. 2014 Feb;32(2):101–8.
14. Guyot P, Welton NJ, Ouwens MJNM, Ades AE. Survival Time Outcomes in Randomized, Controlled Trials and Meta-Analyses: The Parallel Universes of Efficacy and Cost-Effectiveness. Value Health. 2011 Jul;14(5):640–6.
15. Bagul A, Berle S. Survival analysis and extrapolation modelling of time to event clinical trial data for economic evaluation An alternative approach, Medical Decision Making, 343–351
16. Gorrod H, Kearns B, Stevens J. et al. A Review of Survival Analysis Methods Used in NICE Technology Appraisals of Cancer Treatments: Consistency, Limitations, and Areas for Improvement, Medical Decision Making 2019, Vol. 39(8) 899–909
17. Kazandjian, D., Suzman, D., Blumenthal, G. et al Characterization of outcomes in patients with metastatic non-small cell lung cancer treated with programmed cell death protein 1 inhibitors past RECIST version 1.1-defined disease progression in clinical trials. Seminars in Oncology, 10

Feb 2017, 44(1):3-7
18. Karrison, T., Versatile Tests for Comparing Survival Curves Based on Weighted Log-rank Statistics The Stata Journal (2016)16, Number 3, pp. 678–690
19. Chhatwal,J. Jayasuriya, S., Elbasha, E., Changing Cycle Lengths in State-TransitionModels: Challenges and Solution, Med Decis Making 2016;36:952–964
20. Basu, A., and Manca, A., Regression Estimators for Generic Health-Related Quality of Life and Quality-AdjustedLife Years, Med Decis Making 2012;32:56–69
21. Sumner, W Ding, E., Fischer, I., et alMethods for Performing Survival Curve Quality-of-Life Assessments, Med Decis Making 2014;34:787–799)
22. Glover, M., Jones E., Masconi, K., et al, Discrete Event Simulation for Decision Modeling in Health Care: Lessons from Abdominal Aortic Aneurysm Screening, Medical Decision Making 38(4) 439-451