NICE DSU TECHNICAL SUPPORT DOCUMENT 14: SURVIVAL ANALYSIS FOR ECONOMIC EVALUATIONS ALONGSIDE CLINICAL TRIALS - EXTRAPOLATION WITH PATIENT-LEVEL DATA

REPORT BY THE DECISION SUPPORT UNIT

June 2011

(last updated March 2013)

Nicholas Latimer

School of Health and Related Research, University of Sheffield, UK

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street

Sheffield, S1 4DA

Tel (+44) (0)114 222 0734

E-mail dsuadmin@sheffield.ac.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University.

The DSU is commissioned by The National Institute for Health and Clinical Excellence (NICE) to provide a research and training resource to support the Institute's Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

The NICE Guide to the Methods of Technology Appraisal[i] is a regularly updated document that provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The Methods Guide does not provide detailed advice on how to implement and apply the methods it describes. This DSU series of Technical Support Documents (TSDs) is intended to complement the Methods Guide by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in each topic area, and make clear recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE Technology Appraisals, whether manufacturers, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Dr Allan Wailoo

Director of DSU and TSD series editor.

[i] National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal, 2008 (updated June 2008), London.

Acknowledgements

The DSU thanks Tony Ades, Peter Clark, Neil Hawkins, Peter Jones, Steve Palmer and the team at NICE, led by Gabriel Rogers, for reviewing this document.

The production of this document was funded by the National Institute for Health and Clinical Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the author only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

This report should be referenced as follows:

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

EXECUTIVE SUMMARY

Interventions that impact upon survival form a high proportion of the treatments appraised by NICE, and in these it is essential to accurately estimate the survival benefit associated with the new intervention. This is made difficult because survival data is often censored, meaning that extrapolation techniques must be used to obtain estimates of the full survival benefit. Where such analyses are not completed estimates of the survival benefit will be restricted to that observed directly in the relevant clinical trial(s) and this is likely to represent an underestimate of the true survival gain. This leads to underestimates of the Quality Adjusted Life Years gained, and therefore results in inaccurate estimates of cost-effectiveness.

There are a number of methods available for performing extrapolation. Exponential, Weibull, Gompertz, log-logistic or log normal parametric models can be used, as well as more complex and flexible models. The different methods have varying functional forms and are likely to result in different survival estimates, with the differences potentially large – particularly when a substantial amount of extrapolation is required. It is therefore very important to justify the particular extrapolation approach chosen, to demonstrate that extrapolation has been undertaken appropriately and so that decision makers can be confident in the results of the associated economic analysis. Statistical tests can be used to compare alternative models and their relative fit to the observed trial data. This is important, particularly when there is only a small amount of censoring in the dataset and thus the extrapolation required is minimal. However it is of even greater importance to justify the plausibility of the extrapolated portion of the survival model chosen, as this is likely to have a very large influence on the estimated mean survival. This is difficult, but may be achieved through the use of external data sources, biological plausbility, or clinical expert opinion.

A review of the survival analyses included in NICE Technology Appraisals (TAs) of metastatic and/or advanced cancer interventions demonstrates that a wide range of methods have been used. This is to be expected, because different methods will be appropriate in different circumstances and contexts. However the review also clearly demonstrates that in the vast majority of TAs a systematic approach to survival analysis has not been taken, and the extent to which chosen methods have been justified differs markedly between TAs and is usually sub-optimal.

In the form of a Survival Model Selection Process algorithm we provide recommendations for how survival analysis can be undertaken more systematically. This involves fitting and testing a range of survival models and comparing these based upon internal validity (how well they fit to the observed trial data) and external validity (how plausible their extrapolated portions are). Following this process should improve the likelihood that appropriate survival models are chosen, leading to more robust economic evaluations.

CONTENTS

Abbreviations and definitions

AIC	Akaike's Information Criterion
AG	Assessment Group
BIC	Bayesian Information Criterion
DSU	Decision Support Unit
ERG	Evidence Review Group
FAD	Final Appraisal Determination
GIST	Gastro-intestinal Stromal Tumours
HR	Hazard Ratio
ITT	Intention To Treat
LRIG	Liverpool Reviews and Implementation Group
MRC	Medical Research Council
MTA	Multiple Technology Appraisal
NICE	National Institute for Health and Clinical Excellence
OS	Overall Survival
PFS	Progression Free Survival
PH	Proportional Hazards
RCT	Randomised Controlled Trial
SEER	Surveillance, Epidemiology and End Results Program
SMEEP	Survival Model Selection for Economic Evaluations Process Chart
STA	Single Technology Appraisal
TA	Technology Appraisal
TSD	Technical Support Document
US	United States

1. INTRODUCTION

Interventions that impact upon survival form a high proportion of the treatments appraised by the National Institute for Health and Clinical Excellence (NICE). Survival modelling is required so that the survival impact of the new intervention can be taken into account alongside health related quality of life impacts within health economic evaluations. This requirement is reflected by the NICE Guide to the Methods of Technology Appraisal[1], which states that a lifetime time horizon should be adopted in evaluations of interventions that affect survival at a different rate compared to the relevant comparators. Estimates of entire survival distributions are required to ensure that mean impacts on time-to-event (such as progression-free survival and overall survival) are derived, as it is mean rather than median effects that are important for economic evaluations. However, survival data are commonly censored hence standard statistical methods cannot be used, and thus different approaches are required. There are many approaches for conducting survival analysis in these circumstances, and a range of different methods have been used in NICE Technology Appraisals (TAs). However, there is currently no methodological guidance advising when different methods should be used. This leads to the potential for inconsistent analyses, results and decision-making between TAs. The main problem with this is not that different methods are used for estimating survival in different TAs – as different approaches may be appropriate in different circumstances – but rather that the methods used are not justified in a systematic way and often appear to be chosen subjectively. Hence different analysts may select different techniques and models, some of which might be inappropriate, without justification or adequate consideration of the robustness of the model results to alternative approaches.

This Technical Support Document (TSD) provides examples of different survival analysis methodologies used in NICE Appraisals, and offers a process guide demonstrating how survival analysis can be undertaken more systematically, promoting greater consistency between TAs. The focus is on situations in which patient-level data are available, and where evidence synthesis between trials is *not* required – that is, effectively where an economic evaluation is undertaken alongside one key clinical trial. Two other contexts are common in NICE Appraisals – modelling based only upon summary statistics because patient-level data are not available; and modelling where patient-level data are available for one key trial but not for relevant trials of key competitors. The methods for evidence synthesis required under these circumstances are not considered in this TSD – instead an elementary guide for fitting survival models to patient-level data from one trial is presented. It is anticipated that a TSD

addressing survival modelling using summary statistics and evidence synthesis will be produced in the future. Where only summary statistics are available analysts should consider the use of methods introduced by Guyot et al (2012) in order to re-create patient level data.[2] In cases where evidence synthesis is required to include all relevant comparators within a TA analysts should not rely only upon the methods discussed in this TSD.

The TSD is set out as follows. First, a set of standard methods for conducting survival analysis that have been used in past NICE Appraisals, and methods for assessing the suitability of survival models are summarised. Then, the survival analysis methods used in a range of NICE TAs are reviewed, which will serve to highlight potential deficiencies in certain methods and inconsistencies between Appraisals. Finally guidance is put forward to suggest a process by which survival analysis should be carried out in the context of a NICE TA. Importantly, the emphasis within this TSD is upon the process for undertaking survival analysis, rather than exhaustively specifying all potentially relevant methods. The methods discussed in detail here are standard methods documented in many statistics textbooks, and which have regularly been used in NICE TAs. The key focus of this TSD is on those methods which are commonly used in TAs rather than emerging or novel approaches. Further, methods used to account for treatment crossover, which potentially biases mean estimates based upon an intention to treat (ITT) analysis are not considered in this TSD – a separate document is required for that topic. We acknowledge that other more complex and novel methods are available, and while a selection of these are mentioned they are not reviewed in detail. The absence of discussion around specific relevant methods in this TSD does not preclude their use.

Because this TSD focuses on situations where patient-level data are available it is particularly relevant to those preparing sponsor submissions to NICE. However, undertaking and reporting survival analysis as suggested in this TSD will also enable Assessment Groups (AGs) to critique sponsor submissions more effectively and in circumstances where patient-level data are provided to AGs, they should follow the processes outlined here.

2. SURVIVAL ANALYSIS MODELLING METHODS

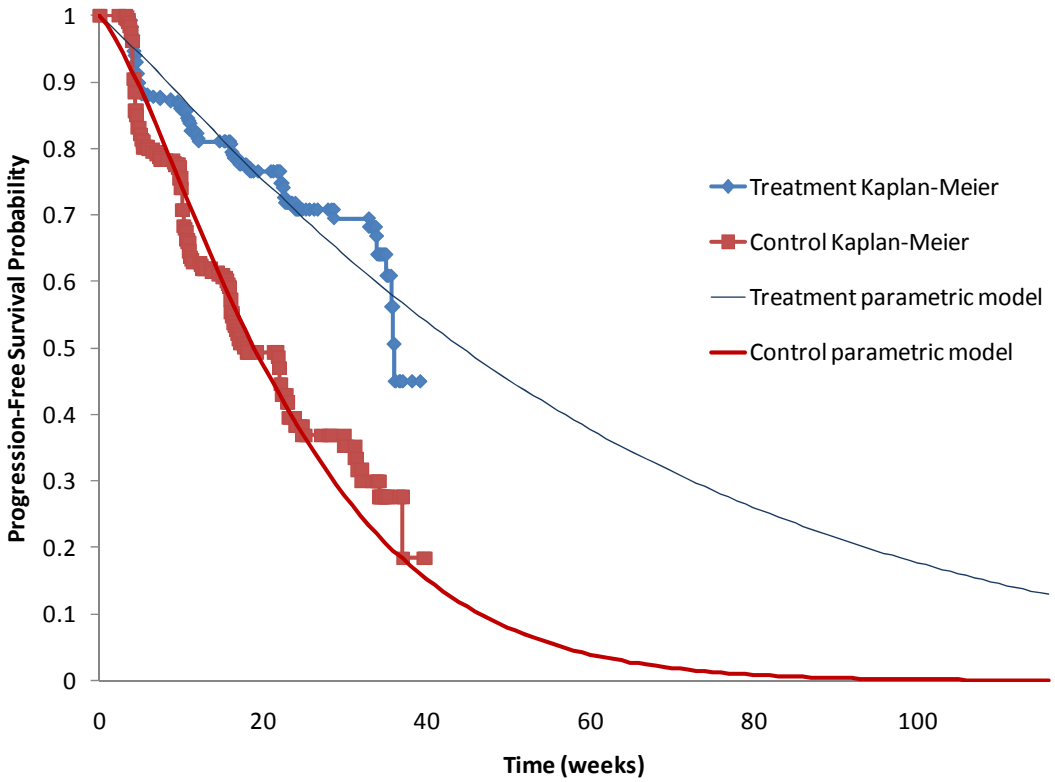
Survival analysis refers to the measurement of time between two events – in clinical trials this is usually the time from randomisation to disease progression (often referred to as progression-free survival, particularly in cancer disease areas) or death (overall survival). Survival data are different from other types of continuous data because the endpoint of interest is often not observed in all subjects – patients may be lost to follow-up, or the event may not have occurred by the end of study follow-up. Data for these patients are censored but are still useful as they provide a lower bound for the actual non-observed survival time for each censored patient.[3] Survival analysis techniques allow these data to be used rather than excluded; however there are a range of different survival distributions and models that can be used. The choice of model can lead to different results. These alternative approaches will be discussed here. It is important to note that the standard survival analysis methods discussed here are only suitable if censoring is uninformative (that is, any censoring is random).

Unless survival data from a clinical trial is complete, or very close to being complete – that is, most patients have experienced the event by the end of follow-up – extrapolation is required such that survival data can be usefully incorporated in health economic models. Generally speaking, this is achieved through the use of parametric models which are fitted to empirical time to event data. Alternatives exist, however these are generally not appropriate when censoring is present. For example a restricted means analysis usually involves estimating the mean based only upon the available data (although it could also mean only extrapolating up to a certain time point). Similarly, a Cox proportional hazards regression model, discussed later, bases inferences only upon observed data. However such methods are only likely to be reasonable when data is almost entirely complete, as otherwise they will not produce accurate estimates of mean survival and will not reflect the full distribution of expected survival times, as these are affected by omitting more extreme extrapolated datapoints. Therefore parametric models are likely to represent the preferred method for incorporating survival data into health economic models in the majority of cases. In this situation the problem becomes one of how to best make inferences about the tails of probability distributions given partial – or even completely absent – information. For example, care must be taken in the common case where lifetime data are immature and non-censored observed values are only available on a small proportion of patients.

Figure 1 illustrates how survival data may be extrapolated using a parametric model. The diagram illustrates the non-parametric Kaplan Meier estimate of the survivor function for the event of interest (in this case progression-free survival) over time for a control group and a treatment group, taken directly from clinical trial survival data. In this example, follow-up ends after approximately 40 weeks, at which point approximately 45% of treatment group patients who had not been censored up until this point had not experienced the event of interest. The equivalent figure is approximately 20% in the control group. Since the chart plots survival over time for the trial population, the mean survival of the trial population is equal to the area under the curve. However, because a proportion of patients remain alive at the end of the 40 week follow-up period only a mean restricted to this time point can be directly estimated. As mentioned above, parametric models can be used to avoid a reliance on restricted mean estimates. Figure 1 shows parametric extrapolations of the survival data (in this case Weibull models have been used) which demonstrates how survival data can be extrapolated so that an unrestricted estimate of mean survival for each treatment group can be obtained – models are fitted and the total area under the curve can be estimated. In these circumstances the base case analysis should use extrapolation of the fitted probability distribution, although also presenting results based only upon the observed data may provide useful information regarding the importance of the extrapolated period in the determination of the mean.

Figure 1 also demonstrates the 'stepped' nature of Kaplan Meier curves, which occurs because follow-up only occurs at pre-specified time intervals – in this instance every 6 weeks. This means that events are only observed to have occurred at 6-week intervals. In some cases this could create bias in survival analysis results – particularly where follow-up intervals are relatively long. In these circumstances interval censoring methodology should be considered. Possible methods are discussed by Collett (2003)[4] and are available in standard statistical software packages. Related issues are discussed by Panageas et al (2007)[5]. The approach taken should be justified with respect to its use in the economic model.

Figure 1: Kaplan Meier curves and parametric extrapolations



There are a wide range of parametric models available, and each have their own characteristics which make them suitable for different data sets. Exponential, Weibull, Gompertz, log-logistic, log normal and Generalised Gamma parametric models should all be considered. These models, and methods to assess which of these models are suitable for particular data sets are described below. Further details on the properties of the individual parametric models that should be considered can be found in Collet (2003)[4], including diagrams of hazard, survivor and probability density functions which show the variety of shapes that the different models can take, depending upon their parameters. The hazard function is the event rate at time *t* conditional upon survival until time *t*. The survivor function is the probability that the survival time is greater than or equal to time *t* and is equivalent to $1 - F(t)$ where $F(t)$ is the probability density function, representing the probability that the survival time is less than *t*.

2.1 EXPONENTIAL DISTRIBUTION

Hazard function: $h(t) = \lambda$ for $0 \leq t < \infty$ where λ is a positive constant and *t* is time.

Survivor function: $S(t) = \exp\left\{-\int_0^t \lambda \mathrm{d}u\right\} = e^{-\lambda t}$

The exponential distribution is the simplest parametric model as it incorporates a hazard function that is constant over time, and therefore it has only one parameter, λ. The exponential model is a proportional hazards model, which means that if two treatment groups are considered within the model, the hazard of the event for an individual in one group at any time point is proportional to the hazard of a similar individual in the other group – the treatment effect is measured as a hazard ratio. Methods for assessing the suitability of alternative parametric distributions and the validity of the proportional hazards assumption will be considered in detail below, but if the exponential distribution is to be used it is important to consider whether the hazard is likely to remain constant over an entire lifetime.

2.2 WEIBULL DISTRIBUTION

Hazard function: $h(t) = \lambda\gamma t^{\gamma-1}$ for $0 \leq t < \infty$ where λ is a positive value and is the scale parameter, and γ is a positive value and is the shape parameter.

Survivor function: $S(t) = \exp\left\{-\int_0^t \lambda\gamma u^{\gamma-1}\mathrm{d}u\right\} = \exp(-\lambda t^{\gamma})$

The Weibull distribution can be parameterised either as a proportional hazards model (as shown in the survivor function above) or an accelerated failure time model. In an accelerated failure time model when two treatment groups are compared the treatment effect is in the form of an acceleration factor which acts multiplicatively on the time scale. Weibull models depend on two parameters – the shape parameter and the scale parameter. The Weibull distribution is more flexible than the exponential because the hazard function can either increase or decrease monotonically, but it cannot change direction. The exponential distribution is a special case of the Weibull, where $\gamma = 1$. Where $\gamma > 1$ the hazard function increases monotonically and where $\gamma < 1$ the hazard function decreases monotonically. When considering the applicability of a Weibull distribution the validity of monotonic hazards must be considered.

2.3 GOMPERTZ DISTRIBUTION

Hazard function: $h(t) = \lambda e^{\theta t}$ for $0 \leq t < \infty$ where λ is a positive value and is the scale parameter, and θ is the shape parameter.

Survivor function: $S(t) = \exp\left\{\frac{\lambda}{\theta}\left(1 - e^{\theta t}\right)\right\}$

Similar to the Weibull distribution the Gompertz has two parameters – a shape parameter and a scale parameter. Also similar to the Weibull distribution the hazard in the Gompertz distribution increases or decreases monotonically. Where θ = 0 survival times have an exponential distribution, where θ > 0 the hazard increases monotonically with time and where θ < 0 the hazard decreases monotonically with time. The Gompertz distribution differs from the Weibull distribution 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. Also, the Gompertz model can only be parameterised as a proportional hazards model. When considering the applicability of a Gompertz distribution the validity of monotonic hazards must be considered.

2.4 LOG-LOGISTIC DISTRIBUTION

Hazard function: $h(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{1+e^{\theta}t^{\kappa}}$ for $0 \leq t < \infty, \kappa > 0$

Survivor function: $S(t) = \left\{1 + e^{\theta}t^{\kappa}\right\}^{-1}$

The log-logistic distribution is an accelerated failure time model and has a hazard function which can be non-monotonic with respect to time. It has two parameters, θ and κ. If $\kappa \leq 1$ the hazard decreases monotonically with time, but if $\kappa > 1$ the hazard has a single mode whereby there is initially an increasing hazard, followed by a decreasing hazard. When considering the applicability of the log-logistic distribution the validity of non-monotonic hazards must be considered. Owing to their functional form, log-logistic models often result in long tails in the survivor function, and this must also be considered if they are to be used.

2.5 LOG NORMAL DISTRIBUTION

Hazard function: $h(t) = \frac{f(t)}{S(t)}$ for $0 \leq t < \infty$ where *f(t)* is the probability density function of *T*.

Survivor function: $S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$ where Φ is the standard normal distribution function.

The log normal distribution is very similar to the log-logistic distribution, and has two parameters: μ and μ. The hazard increases initially to a maximum, before decreasing as *t* increases. 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 non-monotonic hazards must be considered, and the validity of potentially long tails in the survivor function must be considered.

2.6 GENERALISED GAMMA

Hazard function: $h(t) = f(t)/S(t)$ where *f(t)* is the probability density function of *T*.
Survivor function: $S(t) = 1 - \Gamma_{(\lambda t)^{\theta}}(\rho)$ where $\Gamma_{\lambda t}(\rho)$ is known as the incomplete gamma function.

The Generalised Gamma distribution is a flexible three-parameter model, with parameters λ, ρ and θ. It is a generalisation of the two parameter gamma distribution and it is useful because it includes the Weibull, exponential and log normal distributions as special cases, which means it can help distinguish between alternative parametric models. θ is the shape parameter of the distribution and when this equals 1 the generalised gamma distribution is equal to the standard gamma distribution. When ρ equals 1 the distribution is the same as the Weibull distribution and as ρ becomes closer to infinity the distribution becomes more and more similar to the log normal distribution. Hence when a generalised gamma model is fitted the resulting parameter values can signify whether a Weibull, Gamma or log normal model may be suitable for the observed data.

2.7 PIECEWISE MODELS

Piecewise parametric models represent an under-used modelling approach in health technology assessment. These models are more flexible than individual parametric models and provide a simple way for modelling a variable hazard function. They are generally referred to as piecewise constant models, as typically exponential models are fitted to different time periods, with each time period having a constant hazard rate.[6] Piecewise constant models are particularly useful for modelling datasets in which variable hazards are observed over time. Models other than the exponential also allow for non-constant hazards over time, but in the case of Weibull and Gompertz models the hazard must be monotonic, and in the case of log-logistic and log normal models the hazard is unimodal. Piecewise constant models do not restrict the hazard in this way. However, these models are less useful for the extrapolated portion of the survival curve, since in this portion hazards are not

observed. Thus, as an alternative to the piecewise constant model consideration could be given to using a different parametric model (such as a Weibull, Gompertz, log-logistic, log normal of Generalised Gamma model) for the extrapolated portion of the survival curve, although an exponential should also be considered if it is deemed appropriate to extrapolate with a constant hazard rate. Consideration of how external data and information might be used to inform the decision as to which parametric model is most appropriate for long-term extrapolation is given below.

2.8 OTHER MODELS

Alongside the standard parametric models and piecewise models discussed above there are various other more weakly structured, flexible models available – such as Royston and Parmar's spline-based models.[7] These have not been used in NICE Appraisals as yet, but are potentially very useful. They are flexible parametric survival models that resemble generalised linear models with link functions. In simple cases these models can simplify to Weibull, Log-logistic or log normal distributions – which demonstrates their flexibility and usefulness in discriminating between alternative parametric models. Jackson et al (2010) discuss and implement other flexible parametric distributions, such as the Generalised F – which has four parameters and which simplifies to the Generalised Gamma distribution when one of those parameters tends towards zero – as well as Bayesian semi-parametric models which allow an arbitrarily flexible baseline hazard, and which are extrapolated by making assumptions about the future hazard (ideally based upon additional data or expert judgement).[8] These more flexible methods have not been used in NICE Appraisals as yet, but Jackson et al provide a helpful case study of the application of these methods, and the determination of best fitting models.

2.9 MODELLING APPROACHES

When a parametric model is fitted to survival data two broad approaches may be taken. One option is to split the data and fit an individual or piecewise parametric model to each treatment arm. The second option is not to split the data and to fit one parametric model to the entire dataset, with treatment group included as a covariate in the analysis and assuming proportional hazards. The approach taken is very often likely to reflect the nature of the comparison being drawn.

When there are multiple comparators which have been examined in separate RCTs there is often a reliance on summary statistics, which lends itself to a proportional hazards modelling approach using hazard ratios. Under this approach a hazard ratio (HR) is applied to a base survival curve to compare an experimental treatment to a control so that all treatments can be compared to a common comparator. Where one HR is applied to the entire modelled period, the proportional hazards assumption must be made – that is, the treatment effect is proportional over time and the survival curves fitted to each treatment group have a similar shape. The approach can be used within proportional hazards models such as the exponential, Gompertz or Weibull but log-logistic and log normal models are accelerated failure time models and do not produce a single hazard ratio (HR), and thus the proportional hazards assumption does not hold with these models. However, modelling using treatment group as a covariate can still be undertaken with these models, with the treatment effect measured as an 'acceleration factor' rather than a HR.

Generally, when patient-level data are available, it is unnecessary to rely upon the proportional hazards assumption and apply a proportional hazards modelling approach – the assumption should be tested which will indicate whether it may be preferable to separately fit parametric models to each treatment arm, or to allow for time-varying hazard ratios. Fitting separate parametric models to each treatment arm involves fewer assumptions, although it does also require the estimation of more parameters. While fitting separate parametric models to individual treatment arms may be justified, it is important to note that fitting different types of parametric model (for example a Weibull for one treatment arm and a log normal for the other) to different treatment arms would require substantial justification, as different models allow very different shaped distributions. Hence if the proportional hazards assumption does not seem appropriate it is likely to be most sensible to fit separate parametric models of the same type, allowing a two-dimensional treatment effect on both the shape and scale parameters if the parametric distribution.[9]

If a proportional hazards model is used, the proportional hazards assumption and the duration of treatment effect assumption should be justified (using methods described below). In addition, care should be taken to ensure that only the HR obtained from the chosen parametric model is applied to the control group survival curve derived from the parametric model fitted with the treatment group as a covariate – it is theoretically incorrect to apply a HR derived from a different parametric model, or one derived from a Cox proportional

hazards model.[10] There are practical implications of this when modelling is based upon summary data rather than patient-level data, as the origin of quoted HRs may not be clear. It is anticipated that this issue will be considered in a future TSD that considers evidence synthesis and survival analysis.

3. ASSESSING THE SUITABILITY OF SURVIVAL MODELS

There are a variety of methods that can and should be used when assessing the suitability of each fitted model. A range of methods that are likely to be of use are described briefly below. This is not intended to be an exhaustive list, since several other statistical tests may be useful (for example, tests of residuals such as Cox-Snell, Martingale or Schoenfeld residuals), but those listed below are likely to be particularly relevant. Assessing the suitability of alternative survival models is concerned with demonstrating whether or not models are appropriate, which is defined by whether the model provides a good fit to the observed data *and* whether the extrapolated portion is clinically and biologically plausible. Models that meet only one of these criteria are likely to be inappropriate.

3.1 VISUAL INSPECTION

It is often useful to assess how well a parametric survival model fits the clinical trial data by considering how closely it follows the Kaplan Meier curve visually. This provides a simple method by which one model could be chosen over another. However, this method of assessment is uncertain and may be inaccurate. If censoring is heavy and observed data points are clustered at certain points along the Kaplan Meier curve, it might be quite reasonable for a parametric model to follow the Kaplan Meier closely for one segment, but not at another – such an occurrence does not necessarily mean that the model is inappropriate. In addition, a fitted model may follow the Kaplan Meier curve closely but may have an implausible tail (which might be determined through, for example, the use of external data or through clinical expert opinion). Hence the use of this approach for assessing the suitability of parametric models should be used with caution and should be supplemented with other tests.

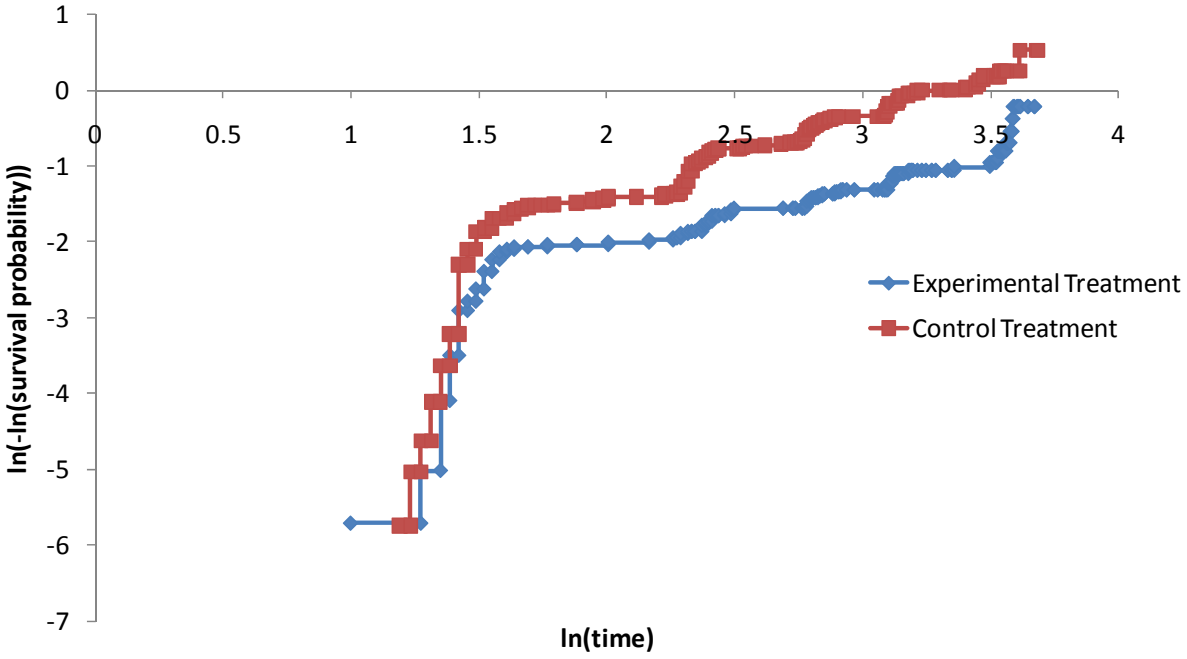
3.2 LOG-CUMULATIVE HAZARD PLOTS

Consideration of the observed hazard rates over time is important when considering suitable parametric models. Different parametric models incorporate different hazard functions. Exponential models are only suitable if the observed hazard is approximately constant and non-zero. Weibull and Gompertz models incorporate monotonic hazards, while the Log-logistic and log normal models can incorporate non-monotonic hazards but typically have long tails due to a reducing hazard as time increases after a certain point. More details are available from various statistical publications, including Collett (2003).[4]

Log-cumulative hazard plots can be constructed to illustrate the hazards observed in the clinical trial. These allow an inspection of whether hazards are likely to be non-monotonic, monotonic or constant. In addition, these plots allow an assessment of whether the proportional hazards assumption – which underpins the proportional hazards modelling technique – is reasonable. The plots also show where significant changes in the observed hazard occur, which can be useful when considering the use of different parametric models for different time periods in a piecewise modelling approach. Standard log-cumulative hazard plots (a plot of: log (-log of the survivor function) against log (time)) are used to test the suitability of the Weibull and exponential distributions. Variations on this approach can be used to test the suitability of the Gompertz, log normal and log-logistic distributions. Again, more details are available in Collett (2003). [4]

Figure 2 shows an illustration of a log cumulative hazard plot for the Kaplan Meier curves previously shown in Figure 1. It demonstrates that there is a seemingly important change in the hazard after approximately 5 weeks (exp(1.5)), but that hazards are reasonably proportional between the two treatment groups. This signals that a single parametric model may not be suitable to model survival, although the hazards observed prior to the 5-week timepoint (and the 'steps' later on in the plots) may at least partially be explained by interval censoring. The gradient of the plot after 5 weeks for the experimental group appears to be less than 1 and hence an exponential model is unlikely to be suitable. After 5 weeks, the gradients of the plots are reasonably constant and so a Weibull model may be suitable after this timepoint, although there is a steepening of the experimental treatment plot towards the end of follow-up that would be worthy of further investigation. Variations of the log-cumulative hazard plot should be used to test the suitability of other parametric models.

Figure 2: Log-cumulative hazard plot



3.3 AIC/BIC TESTS

Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) provide a useful statistical test of the relative fit of alternative parametric models, and they are usually available as outputs from statistical software. Further details on these are available from Collett (2003).[4] Measures such as the negative 2 log likelihood are only suitable for comparing nested models, whereby one model is nested within another (for example, one model adds an additional covariate compared to another model). Different parametric models which use different probability distributions cannot be nested within one another. Thus the negative 2 log likelihood test is not suitable for assessing the fit of alternative parametric models, and it has been used erroneously in past NICE TAs. The AIC and BIC allow a comparison of models that do not have to be nested, including a term which penalises the use of unnecessary covariates (these are penalised more highly by the BIC). Generally it is not necessary to include covariates in survival modelling in the context of an RCT as it would be expected that any important covariates would be balanced through the process of randomisation. However, some parametric models have more parameters than others, and the AIC and BIC take account of these – for example an exponential model only has one parameter and so in comparative terms two-parameter models such as the Weibull or Gompertz models are penalised. The AIC and BIC statistics therefore weigh up the improved

fit of models with the potentially inefficient use of additional parameters, with the use of additional parameters penalised more highly by the BIC relative to the AIC.

3.4 OTHER METHODS

Other methods for internally validating a model that have not been used in NICE Appraisals but which can be helpful include: 1) Splitting the observed data at random, developing a model based upon one portion and evaluating it on another, and 2) k-fold cross validation and bootstrap resampling, as described by Harrell (2001).[11] In addition, the case study reported by Jackson et al (2010) makes use of the deviance information criterion (DIC), which is a generalisation of the AIC/BIC tests, and which the authors use to assess the expected ability of various models to predict beyond the observed data.[8] Methods such as these should be given due consideration when attempting to justify fitted survival models through statistical analyses.

3.5 LIMITATIONS OF THE ABOVE APPROACHES

An important limitation that is applicable to visual inspection, log-cumulative hazard plots and AIC/BIC tests is that each are based only upon the relative fit of parametric models to the observed data. While this is useful as it is important to determine which models fit the observed data best, it does not tell us anything about how suitable a parametric model is for the time period beyond the final trial follow-up. In other words, the tests described above address the internal validity of fitted models, but not their external validity. This is of great importance considering the impact that the extrapolated portion of survival curves generally has on estimates of the mean, and demonstrates that there cannot be a reliance only upon these measures when assessing the suitability of alternative models – indeed the reason why we use parametric models is to estimate the extrapolated portion of the curve. If there is a large amount of clinical trial survival data over a long time period it may be reasonable to assume that a parametric model that fits the data well will also extrapolate the trial data well. Also, when survival data are relatively complete the extrapolated portion may contribute little to the overall mean area under the curve and in this case the log-cumulative hazard plots and AIC/BIC test results may be of particular use. However when the survival data require substantial extrapolation it is important to attempt to validate the predictions made by the fitted models by other means.

3.6 CLINICAL VALIDITY AND EXTERNAL DATA

A potentially useful method for assessing the plausibility of the extrapolated portions of parametric survival models is through the use of external data and/or clinical validity. External data could come from a separate clinical trial in a similar patient group that has a longer follow-up, or from long-term registry data for the relevant patient group. If patient-level data could be obtained from such sources, such that long-term survival could be estimated specifically for the patient population included in the clinical trial of the new intervention, this would represent a strong source of information. Along these lines, Royston, Parmar and Altman (2010) provide methods for externally validating a fitted model using an external dataset.[12]

Without access to patient-level data such information can only be indicative, but this is still preferable to no information at all. For example, if a registry states that 5-year survival for a particular disease is 10%, parametric models that result in 0% survival at 5 years may not be appropriate, and neither may be those that estimate 40% survival at 5 years. More formally, patient-level data from external data sources could be sought so that more accurate long term survival modelling could be completed, or external data could be used to calibrate fitted models to long-term data-points. However, use of any external data requires a balanced consideration of whether any disparities are likely to be due to a poor extrapolation or limitations in the source of external information.

It is likely that long-term external data will only be available for the control treatment, as by definition the experimental intervention is new. Hence external data is likely to be useful for informing the extrapolation of the control treatment, but may be less helpful for estimating survival on the new intervention in the long-term. Hence, clinically valid and justifiable assumptions on issues such as duration of treatment effect are required to extrapolate long-term survival for the experimental treatment. These could be informed by clinical expert opinion and biological plausibility, and such assumptions should be subject to scenario sensitivity analysis.

Identifying longer term survival evidence and/or obtaining expert clinical judgement on expected long-term hazards should be undertaken routinely when substantial extrapolation is required.

3.7 DEALING WITH UNCERTAINTY

In keeping with the NICE Methods Guide,[1] it is important to consider uncertainty in the analysis of survival data when conducting economic evaluation. When patient-level data are available parameter uncertainty can be taken into account using the variance-covariance matrices for the different parametric models. It is important to note that testing the impact of fitting alternative parametric models – and applying different durations of treatment effect – are effectively types of structural sensitivity analysis.

4. REVIEW OF SURVIVAL ANALYSIS METHODS USED IN NICE TAs

All NICE TAs that dealt with advanced and/or metastatic cancer, or that considered all stages of cancer that had been completed as of December 2009 were reviewed to determine the survival analysis methodology used within the economic evaluation section of the TA. All Appraisal documents available on the NICE website were included in the review, including the assessment report developed by the independent Assessment Group (AG) or Evidence Review Group (ERG), sponsor submissions, final appraisal determinations (FADs), appeal documents, Decision Support Unit (DSU) reports, and documents containing updated analyses. The focus of the review was on methods used to model the entire survival distribution (thus allowing estimates of mean survival) and the rationale given for the approach taken, specifically in those situations whereby patient-level data were available.

The models considered in this TSD can only be fully implemented and justified if patient-level data are available, and so for Appraisals where such data were not available it cannot be expected that the survival analysis will be as systematic. In particular, this may be the case in Multiple Technology Appraisals (MTAs) where the assessment group are not given access to patient-level data. However even in these situations, a sponsor submission that includes analyses based upon patient-level data would be expected. Of the 21 TAs included in the review that occurred since NICE introduced their Single Technology Appraisal (STA) process (from TA110 onwards in the table below) 18 were STAs and only 3 were MTAs (these were TAs 118, 121 and 178). Hence it can be concluded that patient-level data would have been used to inform the primary analysis in the majority of the Appraisals included here, and at least some patient-level data-based analysis can be expected in the vast majority.

Methods used are also dictated by the comparisons required within an Appraisal – several of the TAs reviewed here required comparisons to treatments that were not included in the pivotal trial of the novel intervention and therefore evidence synthesis was required. Guidance on the use of survival analysis methods when evidence synthesis is required is beyond the scope of this TSD, but even when this is the case some analysis of trial data is common (for example, to estimate a baseline survival curve, or for estimating a hazard ratio), and as such some assessment of the suitability of fitted models should be made.

45 TAs were included in the review. The included TAs are listed in table 1.

Table 1: NICE Technology Appraisals (TAs) included in the review

TA Number	Title	Disease Stage	Date Issued
TA3	Ovarian cancer - taxanes (replaced by TA55)	Advanced	May 2000
TA6	Breast cancer - taxanes (replaced by TA30)	Advanced	Jun 2000
TA23	Brain cancer - temozolomide	Advanced	Apr 2001
TA25	Pancreatic cancer - gemcitabine	Advanced / Metastatic	May 2001
TA26	Lung cancer - docetaxel, paclitaxel, gemcitabine and vinorelbine (updated by and incorporated into CG24 Lung cancer)	Advanced / Metastatic	Jun 2001
TA28	Ovarian cancer - topotecan (replaced by TA91)	Advanced	Jul 2001
TA29	Leukaemia (lymphocytic) - fludarabine (replaced by TA119)	Advanced	Sep 2001
TA30	Breast cancer - taxanes (review)(replaced by CG81)	Advanced	Sep 2001
TA34	Breast cancer - trastuzumab	Metastatic	Mar 2002
TA33	Colorectal cancer (advanced) - irinotecan, oxaliplatin & raltitrexed (replaced by TA93)	Advanced	Mar 2002
TA37	Lymphoma (follicular non-Hodgkin's) - rituximab (replaced by TA137)	Advanced / Metastatic	Mar 2002
TA45	Ovarian cancer (advanced) - pegylated liposomal doxorubicin hydrochloride (replaced by TA91)	Advanced	Jul 2002
TA50	Leukaemia (chronic myeloid) - imatinib (replaced by TA70)	All stages	Oct 2002
TA54	Breast cancer - vinorelbine (replaced by CG81)	Advanced / Metastatic	Dec 2002
TA55	Ovarian cancer - paclitaxel (review)	Advanced	Jan 2003
TA62	Breast cancer - capecitabine (replaced by CG81)	Advanced / Metastatic	May 2003
TA61	Colorectal cancer - capecitabine and tegafur uracil	Metastatic	May 2003
TA65	Non-Hodgkin's lymphoma - rituximab	Advanced / Metastatic	Sep 2003

TA Number	Title	Disease Stage	Date Issued
TA70	Leukaemia (chronic myeloid) - imatinib	All stages	Oct 2003
TA86	Gastro-intestinal stromal tumours (GIST) - imatinib	Metastatic	Oct 2004
TA91	Ovarian cancer (advanced) - paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan (review)	Advanced	May 2005
TA93	Colorectal cancer (advanced) - irinotecan, oxaliplatin and raltitrexed (review)	Advanced	Aug 2005
TA101	Prostate cancer (hormone-refractory) - docetaxel	Metastatic	Jun 2006
TA105	Colorectal cancer - laparoscopic surgery (review)	All stages	Aug 2006
TA110	Follicular lymphoma - rituximab	Advanced / Metastatic	Sep 2006
TA116	Breast cancer - gemcitabine	Metastatic	Jan 2007
TA118	Colorectal cancer (metastatic) - bevacizumab & cetuximab	Metastatic	Jan 2007
TA119	Leukaemia (lymphocytic) - fludarabine	All stages	Feb 2007
TA121	Glioma (newly diagnosed and high grade) - carmustine implants and temozolomide	Advanced	Jun 2007
TA124	Lung cancer (non-small-cell) - pemetrexed	Advanced / Metastatic	Aug 2007
TA129	Multiple myeloma - bortezomib	Advanced	Oct 2007
TA135	Mesothelioma - pemetrexed disodium	Advanced	Jan 2008
TA137	Lymphoma (follicular non-Hodgkin's) - rituximab	Advanced / Metastatic	Feb 2008
TA145	Head and neck cancer - cetuximab	Advanced	Jun 2008
TA162	Lung cancer (non-small-cell) – erlotinib	Advanced / Metastatic	Nov 2008
TA169	Renal cell carcinoma - sunitinib	Advanced / Metastatic	Mar 2009
TA171	Multiple myeloma - lenalidomide	Advanced	Jun 2009
TA172	Head and neck cancer (squamous cell carcinoma) - cetuximab	Advanced / Metastatic	Jun 2009
TA174	Leukaemia (chronic lymphocytic, first line) - rituximab	Advanced	Jul 2009
TA178	Renal cell carcinoma	Advanced / Metastatic	Aug 2009
TA176	Colorectal cancer (first line) - cetuximab	Metastatic	Aug 2009
TA179	Gastrointestinal stromal tumours - sunitinib	Advanced / Metastatic	Sep 2009
TA181	Lung cancer (non-small cell, first line treatment) - pemetrexed	Advanced / Metastatic	Sep 2009
TA183	Cervical cancer (recurrent) - topotecan	Metastatic	Oct 2009
TA184	Lung cancer (small-cell) - topotecan	Advanced	Nov 2009

4.1 MODELLING METHODS

The requirement that mean estimates are used for survival parameters (as well as other parameters, such as costs, and health related quality of life) within economic evaluations was generally reflected by the reviewed NICE TAs, with mean time-to-event estimates being used in 36 (80%) of the 45 Appraisals. However median time-to-event estimates were used in 16 of the 45 TAs; this problem was more common in early TAs. Both TAs (TAs 23 and 26) that relied solely upon median survival estimates as parameters within the economic evaluation were completed in 2001. However, even in more recent TAs, some analyses still used median measures of survival times directly within the economic model, either in manufacturer submissions, sensitivity analysis, or where it was deemed that insufficient data was available to reliably estimate a mean. Table 2 summarises the use of means and medians in the reviewed TAs. Here the use of medians relates to using a median directly as a measure of survival in the economic model. Occasionally medians were used so that survival curves could be estimated using an exponential model when patient-level data were not available, and then the area under the exponential model was used within the economic model. This is a reasonable approach when patient-level data are not available, whereas the direct use of a median as a measure of survival in an economic model is not.

Table 2: The use of mean and median survival estimates in NICE Technology Appraisals

Measure	Number of TAs (%)
Means used for any part of the analysis	36 (80%)
Medians used for any part of the analysis	16 (36%)
Means exclusively used (no use of medians for any parameters)	23 (51%)
Medians exclusively used (no use of means for any parameters)	2 (4%)
Unclear which measure was used	5 (11%)

Of the TAs included in the review, the most recent use of median statistics in the survival analysis was in TA171 (Lenalidomide for Multiple Myeloma, completed in June 2009) where medians were used as the point of reference for a calibration exercise using external MRC trial data.[13] The manufacturer argued that calibrating to median survival was preferable because calibrating to the mean would place too great a reliance on unknown event times at the tail of the modelled survival distribution .[14] In contrast, the AG argued that the mean was preferable for economic evaluations, and also noted that in the MRC trials there was very

little censoring and 94% of patients were said to have died – suggesting that there was a relatively small amount of 'unknown' data, and thus the mean estimate was likely to be robust.[15] In another recent TA (TA135, completed in January 2008) the manufacturer argued unsuccessfully that median time-to-event estimates should have been used in the economic evaluation because estimating means involved extrapolation that created uncertainty in the economic model.[16,17] This neglects the fact that health economic models are built to characterise the decision problem and uncertainty – and mean estimates are required to address the decision problem.

In general the use of medians has reduced over time, and when they have been used in recent times usually the AG has criticised this, or it has been due to a lack of patient-level data and with an acknowledgement that mean data is preferable to median data for economic models (eg TAs 23, 26, 54, 62, 119, 121, 135 and 162).

Five broad methods used to estimate mean survival in the reviewed NICE TAs were identified: 1) restricted means analysis; 2) parametric modelling; 3) proportional Hazards (PH) modelling; 4) external data modelling; 5) other 'hybrid' methods. The prevalence of these methods is illustrated in table 3.

Table 3: Methods for estimating mean survival estimates in NICE Technology Appraisals

Method for Estimating Mean	**Number of TAs (%)**
Restricted Means	17 (38%)
Parametric Models	32(71%)
Weibull	23 (51%)
Exponential	20 (44%)
Gompertz	6 (13%)
Log-logistic	9 (20%)
Log normal	6 (13%)
Gamma	2 (4%)
Piecewise modelling	1 (2%)
Proportional Hazards modelling	19 (42%)
External data	4 (9%)
Other 'hybrid' methods	2 (4%)

Method for Estimating Mean	Number of TAs (%)
LRIG Exponential method	1 (2%)
Gelber method	1 (2%)

Parametric models were most commonly used, appearing in 32 (71%) of the 45 TAs. PH modelling (which generally also involves parametric modelling) and restricted means analyses were also common.

4.1.1 Restricted Means Analysis

17 (38%) TAs used a restricted means analysis either for the base case analysis or as a sensitivity analysis. The method as used in the NICE TAs generally involved simply using all the available data to estimate the area under the Kaplan Meier curve up until the final observation, similar to an approach presented in the statistical literature by Moeschberger and Klein (1985).[18] Generally a restricted means approach was only taken when trial data was relatively complete compared to situations where parametric modelling was used.

4.1.2 Parametric Modelling

The majority of TAs (32 (71%) of the 45 reviewed) used parametric extrapolation techniques in order to produce estimates of survival. The most popular parametric models were the Weibull and exponential – the Weibull being used in 23 (51%) TAs, and the exponential in 20 (44%). An exponential model was often used when Markov models were developed and transition probabilities were not time dependent, and where in the absence of patient-level data analysts transformed median statistics into mean estimates under an exponential assumption (this represents a reasonable use of medians where evidence is lacking, unlike the direct use of median survival times in the economic evaluation, as discussed above). Other models were used much less often, with the Gompertz used in 6 (13%) Appraisals, the log normal used in 6 (13%) Appraisals, the log-logistic used in 9 (20%) Appraisals, and the gamma model used in 2 (4%) Appraisals. In 17 (38%) TAs more than one parametric model was fitted in order to test the fit of different distributions; however this was not done in a systematic way, tests of fit were diverse, and often only two alternative models were tested.

The methods of fitting the parametric models varied. Usually the manufacturer had access to patient-level data and thus could fit parametric models using this, whereas the AG typically had to use a digitising computer program in order to digitally scan published Kaplan Meier curves so that patient-level data could be estimated allowing parametric models to be fitted. This approach is made simpler if data reporting the number of patients at risk data over time are provided alongside Kaplan Meier curves – a practice which is relatively rare but which should be encouraged.

It was most common for all trial data to be used when fitting parametric models. However, in some TAs models were fitted using a restricted data set. For example, in TA86 (imatinib for GIST) and TA121 (carmustine implants and temozolomide for glioma) the sponsor and AG fitted exponential parametric models to trial data up to certain specified time-points.[19,20] The final observed data points were not included due to heavy censoring and associated high levels of uncertainty regarding the observed data at the tail of the distribution. The sponsor and AG suspected that including these data points may allow them to exert undue influence on the parametric model. However, the robustness of such an approach is highly questionable because excluding data points means that the level of uncertainty is increased further. In TA86 the NICE DSU also performed an analysis, and instead of using a restricted data set in line with the AG and manufacturer they included all data points in their model fitting process.[21] A variation on the approach of restricting the data to a certain time-point when fitting parametric models was used in TA169 (sunitinib for renal cell carcinoma) and TA179 (sunitinib for GIST).[22,23] In both cases the AG approved of an approach whereby a Weibull model was fitted to the survival data using only one data point per month. This approach was taken as it allowed the fitted models to follow the Kaplan Meier data more closely from a visual perspective. However this approach implicitly places greater than proportionate weight to segments of the Kaplan Meier where there are fewer data points, and does not place proportionate weight on areas where a large number of data points were observed. Furthermore, this approach requires single data points to be chosen for inclusion in the analysis; the choice of the included points is likely to be arbitrary, whilst excluding other data points leads to greater uncertainty. This is therefore a potentially biased technique, and is directly at odds with the method of excluding data from the right-hand-side of the Kaplan Meier from the analysis – the latter places no weight on the events observed at the right-hand-side of the Kaplan Meier, whereas the former implicitly places a high weight on these events. Both methods are inadvisable.

An alternative model fitting approach was taken in TA121 (carmustine implants and temozolomide for glioma), in which the AG fitted two separate parametric models to two sections of the temozolomide PFS data.[24] One model was fitted to the first 12 months of data, and a second model was fitted to the second 12 months. However, the precise methods used for this piecewise approach are not reported in any of the Appraisal documents.

Although the use of more than one parametric model in 38% of the reviewed TAs and the testing of alternative methods for fitting models suggests that structural uncertainty (that is, uncertainty around the type of parametric model fitted) was addressed to some extent, it is clear that this was not dealt with consistently or systematically.

4.1.3 PH Modelling

Some use of Proportional Hazards (PH) modelling was evident in 19 of the 32 TAs that involved extrapolation of survival data. This involved a baseline parametric survival curve being fitted for the control group and a HR being applied to this to estimate time-to-event for the intervention. Sometimes PH modelling was tested as a structural uncertainty sensitivity analysis (with individual model fitting forming the base case), while in other TAs it was the only method for extrapolation used.

PH modelling was most often used when multiple comparators were included in the evaluation, and where patient-level data were not available for all comparators. For example, in TA70 (imatinib for leukaemia), TA91 (paclitaxel, pegylated liposomal doxorubicin hydrochloride and topotecan for ovarian cancer) and TA93 (irinotecan, oxaliplatin and raltitrexed for colorectal cancer) interventions were indirectly compared by applying a HR for each experimental treatment to a baseline survival curve for a common comparator.[25,26,27] It is anticipated that the use of such a technique will be covered in a future TSD.

However, some use of PH modelling was also made when single comparators were included in the economic model, based mainly upon a single head-to-head RCT for which patient-level data were available. This was the case in TA137 (rituximab for lymphoma) and TA174 (rituximab for leukaemia) in which the manufacturer fitted a Weibull model with a single shape parameter to the control group and intervention group data, which is equivalent to a PH

modelling approach.[28,29] In TA179 (sunitinib for GIST) the manufacturer tested fitting individual models to each treatment arm, as well as the PH modelling technique,[30] and concluded that the PH approach led to curves that did not fit well for the new intervention, based upon a visual inspection. This was therefore left as a sensitivity analysis, with individual parametric models fitted for the base case.

Of the TAs that included PH modelling, relatively few made explicit assumptions about the duration of treatment effect. This is important because without assuming a specific duration of treatment effect it is implicitly assumed that the HR observed in the trial lasts for the entire duration of the economic model – typically a lifetime. Such an assumption may not be reasonable, but only a minority of TAs that used a PH modelling approach explicitly addressed this issue. In TA65 it was assumed that the duration of treatment effect was maintained until the end of the trial follow-up as evidence was available up until this point.[31] In TA70 (imatinib for leukaemia) the AG assumed that the treatment effect was maintained for the duration of time that a patient remained in the chronic phase of the disease, whereas the manufacturer assumed that the effect disappeared after 1 year.[25] In TA137 (rituximab for lymphoma), the manufacturer assumed that the treatment effect was maintained for 5 years,[28] which concerned the AG because a high proportion of lymphoma patients received post-progression treatments, and the impact of a new treatment on the benefits of previous treatments was unknown.[32] The AG suggested that an alternative method might be to assume that the treatment benefit is maintained only until the next treatment is taken. In TA174 (rituximab for leukaemia) the manufacturer assumed that the treatment effect remained until disease progression, based upon post-progression Kaplan Meier curves for the new intervention and the control treatment that were very close together and regularly crossed.[29] This concerned the AG because it involved implicitly assuming an OS benefit that had not been demonstrated by the clinical trial, hence they tested a scenario whereby there was zero OS benefit.[29] Overall, it can be seen that assumptions around the duration of treatment effect differed significantly between TAs.

Importantly, the source model for the HR used in the analysis was only specified in one of the reviewed TAs (TA70). Therefore it is uncertain whether the correct HR was used in the other analyses. If a PH modelling approach is to be taken it should be ensured that a suitable HR is applied – that is, the HR calculated from the parametric model used to model survival with treatment group included as a covariate.

4.1.4 External data

As demonstrated above, in the reviewed TAs survival was typically extrapolated using individual parametric models or PH modelling techniques applied to data from pivotal clinical trials. However, in 4 TAs external registry data were used in the extrapolation of survival estimates, due to a lack of long-term survival data within the trial itself. When external data were used to model long-term survival it was usually either assumed that the risk of death is the same in the post-trial period whether the patient was initially randomised to the intervention or the control treatment, or a PH modelling approach was taken.

In TA110 (rituximab for follicular lymphoma) the manufacturer used trial data to fit a parametric survival model for PFS, but used patient-level data from a large registry to model OS because trial data were very incomplete (median survival had not been reached).[33] The manufacturer fitted an exponential model to the registry data, and applied the same risk of death for all patients once disease progression had occurred, irrespective of their initial randomised treatment group. Thus no additional OS benefits associated with the new treatment were assumed after disease progression, but an OS gain *was* implied because PFS was extended and the risk of death was the same after disease progression. The Assessment Group noted this and stated that although a relationship between PFS and OS had not been proven the manufacturer's analysis implied that 79% of the gain in PFS was translated into an OS gain.[33] In addition, the AG noted that the manufacturer had paid no attention to the similarity or otherwise of the patient population included in the clinical trial, and that included in the registry. They were therefore concerned about the applicability of the registry data, and conducted sensitivity analysis assuming none of the PFS gain was translated to OS, which resulted in an ICER which was still below £20,000 per QALY gained. This gave the Appraisal Committee greater confidence when making their recommendation.[34]

In TA65 (rituximab for non-hodgkin's lymphoma) the manufacturer and the AG both used external patient-level data from a registry to estimate long-term PFS and OS by response category for the control group. The HRs from the relevant clinical trial were then applied to estimate long-term survival for the new intervention.[31] A similar use of external data and PH modelling was used in TA129 (bortezomib for multiple myeloma) due to the short follow-up time of the key clinical trial. The manufacturer used external observational data to estimate survival for the baseline group[35] and then applied HRs for PFS and OS to this base, assuming

that the treatment effect was maintained for 3 years, with the treatment effect reduced after the first year. The AG stated that no rationale for the assumptions around the duration and decline of the treatment effect was given by the sponsor.[36]

External data were used to inform the survival modelling in a slightly different way in TA135 (pemetrexed for mesothelioma). The AG referred to statistics from the Surveillance, Epidemiology and End Results (SEER) Program, a source of cancer statistics from the US to help determine which parametric model might be reasonable for OS.[37] The registry showed that a small proportion of long-term survivors could be expected and as a result the AG rejected an exponential model in favour of a Weibull.

4.1.5 Other 'Hybrid' Methods

Most TAs implemented fairly standard methods when fitting models to estimate mean survival, as described above. Restricted means analyses, individually fitted parametric models, PH modelling and external data have all been used. However some novel approaches have also been used, notably the LRIG Exponential method, and the Gelber method. These methods both involve combining non-parametric (based on the observed data) and parametric analyses (for the extrapolated period).

- *LRIG Exponential*

The Liverpool Reviews and Implementation Group (LRIG) were the AG for TA181 which appraised Pemetrexed for Lung cancer. LRIG stated that the extrapolation techniques used by the manufacturer (exponential and Weibull models) provided poorly fitting survival curves.[38] LRIG obtained patient-level data and examined the cumulative hazard function for each group modelled. They observed that parametric models such as the Weibull, exponential and log normal were not compatible with the trial data across the whole range of observations, which they expected given that hazard rates are unlikely to be proportional and treatment effects may be relatively short-term. They also observed that for each group at some point following the end of treatment the cumulative hazard function assumed a steady linear increase that was indicative of a constant risk of death per unit of time. LRIG stated that the implication of this was that the Kaplan Meier curve itself may be the most appropriate measure of short term time-to-event, but that in the long term the disease progression pathway was likely to resume, which could be modelled using an exponential

distribution. Thus LRIG estimated the area under the Kaplan Meier curve and combined this with the area under an exponential model fitted to the tail of the Kaplan Meier to calculate total mean survival.

- *Gelber method*

A method similar to that developed by LRIG was claimed to have been used in TA118 (bevacizumab and cetuximab for colorectal cancer). The manufacturer of cetuximab stated that they used a method developed by Gelber, Goldhirsch and Cole (1993),[39] which involved fitting an appropriate model to the tail of the OS Kaplan Meier curve and using the estimated model combined with the Kaplan Meier to produce an estimate of total mean OS.[40] Gelber et al state that this method is of particular use when it is easier to fit an appropriate parametric model to the tail of a Kaplan Meier rather than to the Kaplan Meier as a whole.[41] In the method proposed by Gelber log-cumulative hazard probability plots are used to determine the appropriate parametric model and to determine the appropriate values for the point at which the parametric curve takes over from the Kaplan Meier. Both the Gelber method and the LRIG Exponential method are likely to be sensitive to the point at which the parametric model takes over from the Kaplan Meier and therefore if either of these methods are used it is important to provide clear rationale for the switch point using statistical analysis.

However, it is not clear whether the manufacturer in TA118 implemented the Gelber method as the authors would have intended. Parametric curves were fitted from the point at which the OS Kaplan Meier curves for the intervention and the control treatment started to diverge, but no consideration of log-cumulative hazard plots to determine suitable parametric models or the point at which the parametric curve takes over from the Kaplan Meier was reported. The AG noted that the methods used by the manufacturer led to an estimated OS curve that diverged substantially from the Kaplan Meier after around 9 months, with estimated OS seemingly much higher than would be expected.[40] The extrapolated curve flattened after around 9 months, before beginning to fall again, which in no way represented the Kaplan Meier. In the example originally presented by Gelber et al this was not seen, with the extrapolated curve following the Kaplan Meier closely.[41] The AG were therefore concerned that the OS estimated by the manufacturer was unreliable. It seems likely that the manufacturer either implemented the method incorrectly, chose an unsuitable parametric model, or an unsuitable cut-off point.

4.2 MODEL SELECTION

In 22 (69%) of the 32 TAs that used a parametric model-based extrapolation technique, some justification for the extrapolation technique was provided. However the justification given was often very brief and in no TA was a full justification given including an assessment of the fit of all available parametric models, details on the statistical fit of alternative models, tests of proportional hazards, consideration of the expected hazard over time, and a comparison to external or registry data.

The range of methods used for justification purposes in the reviewed TAs are presented in table 4.

Table 4: Methods used to justify the chosen parametric model in NICE Technology Appraisals

Method of Justification	Prevalence in NICE TAs
Statistical tests	
AIC test BIC test Sum of squared deviations -2 log likelihood statistic Log cumulative hazard plot Other tests of the hazard function Martingale residuals	Relatively rare and not systematically done in combination with other methods of justification.
Visual inspection	Common, but often considering only one or a limited subset of possible models.
External data	Rare
Clinical validity	Rare

4.3 VISUAL INSPECTION

The only justification for the chosen parametric model given in a number of TAs was based upon visual inspection. Often if the parametric model was seen to follow the Kaplan Meier curve reasonably closely it was accepted as a 'good fit' and no further justification was given or sought. Regularly the visual inspection would consist only of assessing the one parametric model fitted to the data. However, on a number of occasions one model was classified as a better fit than others based only upon visual inspection.

4.4 Statistical Tests

A range of statistical tests, plots and analyses were used to justify parametric model choices in the different TAs. Tests such as the AIC and BIC were relatively common, while tests of residuals were also occasionally undertaken. Log-cumulative hazard plots were underused given their value in determining which parametric models might be reasonable for a given dataset and their usefulness in determining the suitability of the proportional hazards assumption, which is essential when a PH modelling approach is taken. Testing the internal validity of fitted parametric models by any means other than visual inspection was relatively rare in the reviewed TAs.

4.5 Clinical Validity and External Data

Explicitly testing the external validity of fitted parametric models was even more rare than the testing of internal validity. As discussed above, only TA135 (pemetrexed for mesothelioma) used external data to determine the appropriateness of alternative parametric models. The AG used SEER cancer statistics to demonstrate that a Weibull model was more appropriate than an exponential for modelling OS.[37] However, the Group also noted that many confounding factors may have influenced the SEER statistics, noting the potential limitations associated with using summary statistics from registry data to inform long-term survival estimates. Clinical validity is not a standalone method for justifying model choices, but it should be used alongside other methods.

4.6 Systematic Assessment

Although none of the reviewed TAs included full and systematic justifications for the chosen parametric models, in some TAs multiple justifications were used, but in all cases flaws remained. For example in TA135 (pemetrexed for mesothelioma) the choice of parametric model was based on an analysis of the observed hazard as well as a consideration of expected long-term survival based upon external data, but the exploratory analysis of the hazard was not described and a limited range of parametric models were considered[37] – thus other more suitable models may potentially have been overlooked. In the manufacturer's submission for TA137 (rituximab for lymphoma) consideration was given to a range of parametric models, AIC and BIC tests were conducted and the validity of the extrapolated tails of the survival curves was also considered.[28] However log-cumulative hazard plots were not presented even though a PH modelling approach was taken, and no data were used to justify assumptions

about the expected long-term survival times and thus the suitability of alternative models. Similarly, in TA145 (cetuximab for head and neck cancer) and in TA178 (renal cell carcinoma), a number of parametric models were compared using visual inspection and statistical tests as well as a consideration of the clinical validity of the overall shape of the survival curves. However log-cumulative hazard plots were not constructed, and neither provided any justification to back up the clinical validity assumptions made.[22,42]

In TA172 (cetuximab for head and neck cancer), the manufacturer based their model choice on the log likelihood of the fitted models as well as the clinical validity of the models.[43] However, this approach was flawed because the log likelihood is a test meant for nested models rather than different parametric models, and the Weibull model was the only non-logged model included in the analysis – Gompertz and exponential models were not considered. Log-logistic and log normal models were rejected as their long tails were deemed to be clinically infeasible, but other potentially useful models were overlooked. In TA174 (rituximab for leukaemia) the manufacturer analysed a good range of parametric models – Weibull, exponential, Gompertz, log-logistic and log normal models were considered – and a reasonable range of tests were also conducted – AIC, BIC and martingale residuals.[29] However, log-cumulative hazard plots were not constructed despite the adoption of a PH modelling approach, and little consideration was given to the validity of the models in the extrapolated portion of the curves. In TA184 (topotecan for lung cancer) the AG conducted a range of diagnostic tests (R^2, sum of residuals, log-cumulative hazard plots and visual inspection), but the AIC and BIC were not considered, long-term clinical validity was not considered in detail (although this was a minor issue as little extrapolation was required) and only the Weibull and log-logistic models were compared.[44]

5. REVIEW CONCLUSIONS

It is clear that survival analysis methods have differed significantly in NICE TAs of metastatic and/or advanced cancer interventions. To some extent this is to be expected, because different methods will be appropriate in different circumstances and contexts. However, most importantly, the vast majority of TAs have not taken a systematic approach to survival analysis, and the extent to which chosen methods have been justified differs markedly between TAs. From the review, it is clear that several clarifications are required in order to ensure that survival analysis using patient-level data is conducted more appropriately

in future TAs. These are listed below, and in the following section methodological process guidance is given.

1. Mean time-to-event should be estimated rather than medians.

2. Parametric models should be used, rather than restricted means approaches, unless data is almost entirely complete.

3. The analyst should demonstrate that a range of parametric models have been considered and compared, in order to make evident that the model choice has not been arbitrary. Exponential, Weibull, Gompertz, Log-logistic, log normal and Generalised Gamma models should be considered and if these appear unsuitable due to poor fit or implausible extrapolation, the use of piecewise modelling and other novel survival modelling methods such as those demonstrated by Royston and Parmar and Jackson et al should be considered.[7,8] Where piecewise models are used appropriate distributions should be used for the extrapolated portion.

4. The fit of alternative models should be assessed systematically. Log-cumulative hazard plots (or suitable residuals plots), AIC/BIC tests (or other suitable tests of internal validity), and clinical plausibility based upon expert judgement, external data, or biological reasoning should be presented and assessed. Visual inspection should not be relied upon, but where it is used it is important to include numbers at risk data in diagrams of Kaplan Meier curves, as this aids the review of model fit via visual inspection.

5. PH modelling should only be used if the proportional hazards assumption can be clearly justified using log-cumulative hazard plots, external information and clinical expert opinion. If an PH modelling approach is used the source of the HR used should be clearly stated, and should be taken from the parametric model fitted to the survival data with treatment group included as a covariate.

6. Where parametric models are fitted separately to individual treatment arms it is sensible to use the same 'type' of model, that is if a Weibull model is fitted to one treatment arm a Weibull should also be fitted to the other treatment arm. This allows a two-dimensional treatment effect in that the shape and scale parameters can both differ between

treatment arms, but does not allow the modelled survival for each treatment arm to follow drastically different distributions.[9] If different types of model seem appropriate for each treatment arm this should be justified using clinical expert judgement, biological plausibility, and robust statistical analysis.

7. The duration of treatment effect assumption is important when a PH approach is taken, and in the extrapolated portion of survival curves when individual parametric models are fitted to treatment arms. It is difficult to obtain information on how long the effect of a new treatment may last, but an analysis of the hazards observed in the trial period, clinical expert opinion and biological plausibility should be considered in order to assess the validity of extrapolated curves. As a minimum, undertaking scenarios which match the current NICE Methods Guide should be included – that is, assuming the treatment effect halts at the end of the trial; that it declines over time; and that it is maintained over the lifetime. Scenario based sensitivity analysis should assess the importance of duration of treatment effect assumptions.

8. The approach of excluding data points should only be undertaken when it can be clearly demonstrated that certain points are erroneous outliers. Such evidence might include external data and clinical expert opinion, specifically addressing the validity of the right-hand-side of the Kaplan Meier curves. Instead of excluding data points, a piecewise modelling approach should be considered. Model fitting using one data-point per month helps ensure that the fitted model fits the right-hand-side of the Kaplan Meier, but this approach is likely to be arbitrary, risks over-interpreting this section of the curve, and involves excluding datapoints which is likely to increase uncertainty. Excluding the last observed events prevents these from impacting upon the fit of the model at all and so avoids any risk of over-interpreting this part of the data, but also means that the trial data on longer term survival is ignored and increases uncertainty. Given that data points at the right-hand-side of the Kaplan Meier curve can be particularly influential when fitting parametric models, the different approaches are very likely to lead to significantly different survival estimates. Unless a very clear rationale is offered, *all* data should be included in the survival analysis.

9. External data should be identified using more systematic approaches and used to help inform long-term survival estimates and assessments of external validity of fitted models – either to inform parametric models via techniques such as calibration, to inform assumptions within parametric models, or to directly obtain long-term survival probabilities. Ideally,

patient-level data from relevant external data sources should be obtained so that regression analysis can be completed to allow survival to be estimated adjusting for the characteristics of the patients included in the clinical trial of interest – thus correcting for any patient population differences which may be present between different clinical trials. However, whichever way external data are used, it should be carefully justified, particularly with respect to the patient population.

10. Other approaches observed in previous NICE TAs can be useful. For example the LRIG exponential approach and the Gelber approach are similar to a piecewise exponential method.[38,39,41] The three approaches would be expected to give similar results, but the LRIG exponential and Gelber approaches also partially deal with the problem of long-term extrapolation, by fitting a parametric model only to the tail of the trial data. LRIG suggest that in the long-term the hazard rate may converge to a constant rate, allowing an exponential model to be fitted to the tail of the data. This may or may not be true in other disease areas, but other parametric models with non-constant hazard rates should also be considered, as suggested by Gelber et al (1993). The decision of what model to fit to the tail should be informed where possible by log-cumulative hazard plots and external information.

11. Whatever approach is taken should be systematically justified in comparison to alternative approaches and assumptions, and the robustness of results to these alternatives should be considered. Both parameter and structural uncertainty should be addressed. Where censoring is substantial, one scenario that could be reported is the cost-effectiveness based only upon observed data, as this could provide useful information on the influence of the extrapolated survival period on the incremental cost effectiveness ratio. Often in TAs patient-level data is not available to the AG, making justification of parametric model choices more difficult. However, these data typically are available to manufacturers and the onus is therefore on them to present data and analysis in such a way as to convince the AG and Appraisal Committee that an appropriate survival analysis process has been undertaken, maximising the probability that suitable survival estimates have been obtained. Such confidence could be instilled by manufacturers following the process guide described in the next section.

6. METHODOLOGICAL AND PROCESS GUIDANCE

The survival model selection process is complex and inevitably different models will be appropriate in different TAs. Therefore it is difficult and inappropriate to provide guidance on what methods are 'optimal'. However, it is possible to recommend a process by which model selection can be undertaken, in order to promote process transparency and consistency between TAs. Hence, below we present a model selection process algorithm. Owing to the complexity of the survival modelling process, it is useful for analysts to provide an overview of the process they undertook in order to demonstrate that a logical process was followed. We also present a model selection process chart that could be completed by analysts to make clear the methods used to select a preferred model. Therefore manufacturers and analysts may use the model selection process algorithm (presented in figure 3) for guidance regarding how a preferred model can be selected, and should complete a model selection process chart (example presented in figure 4) to demonstrate the steps they completed.

6.1 MODEL SELECTION PROCESS ALGORITHM

In Figure 3, below, a model selection algorithm is presented which is intended to increase the transparency and consistency in survival analysis methods used in NICE Appraisals when survival models are being fitted to patient-level data in the context of an economic evaluation alongside a key clinical trial in which all relevant comparators are included. This is explained below as a step-by-step process.

Step 1. Log-cumulative hazard plots (or suitable residual plots) should be produced to assess the type of hazards observed in the clinical trial. This helps to demonstrate which type of parametric model is suitable, and whether proportional hazards can be assumed. Log-cumulative hazard plots also highlight situations whereby no single parametric model is suitable to model the observed data.

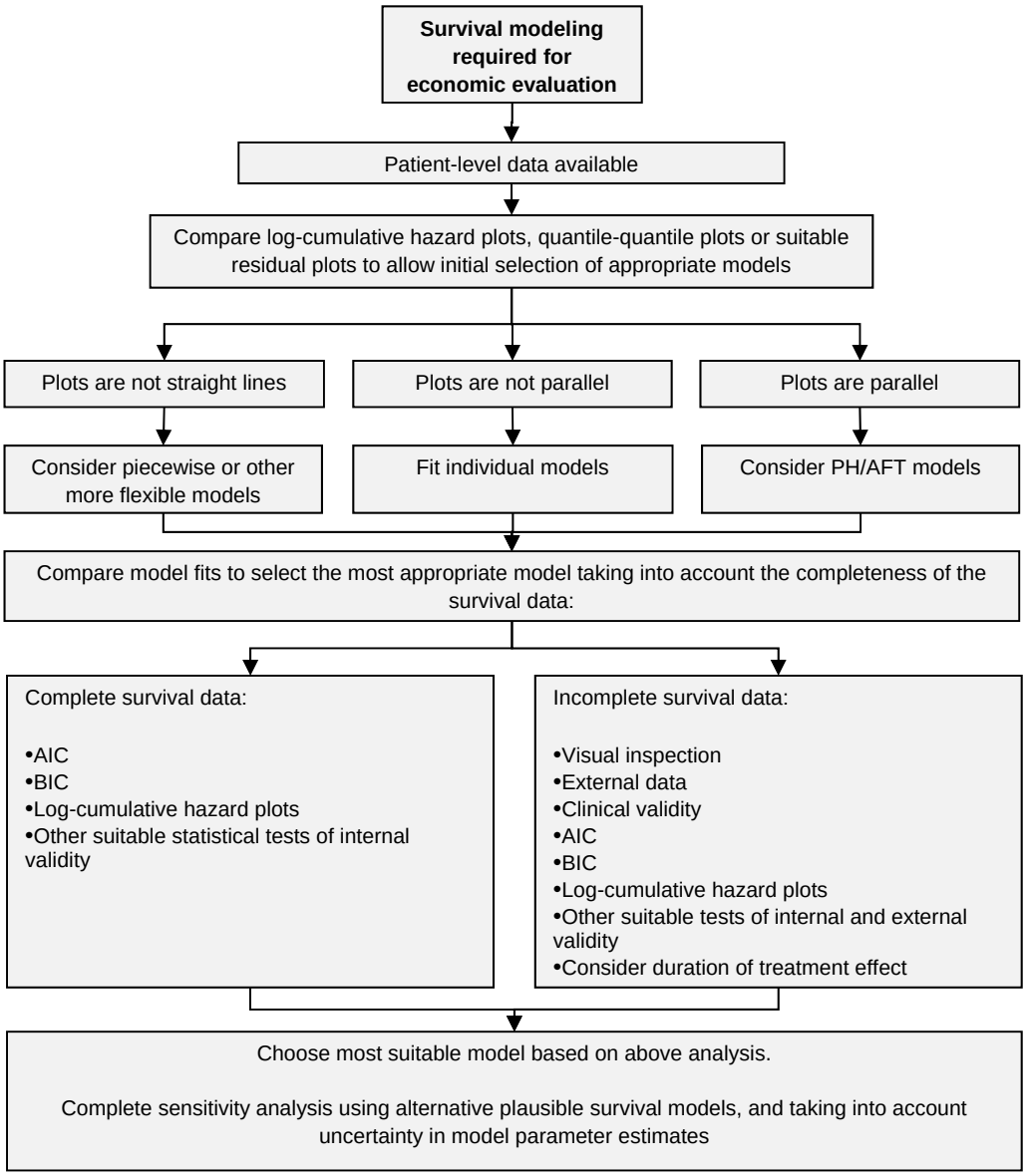
Step 2. If the log-cumulative hazard plots produce approximately straight lines for any of the parametric models then those models should be fitted to the data and assessed further (see later steps). If the plots for the two treatment groups are parallel, proportional hazards models should be considered and assessed further, whereas if they are not parallel, individual model fitting for each treatment arm should be undertaken using a suitable model and assessed further. If the log-cumulative hazard plots are not approximately straight lines,

alternative modelling methods – such as allowing time-varying hazards, piecewise modelling, or more flexible modelling approaches such as those discussed by Jackson et al (2010) or Royston and Parmar (2002) should be considered.[7,8] Visual inspection of the alternative models can be used for assessment purposes, but this technique can be misleading and should not be relied upon as a standalone method.

Step 3. Models deemed potentially appropriate in Step 2 should be compared further using AIC/BIC (or other suitable) tests of internal validity. If data are very close to being complete, model choice can be made based upon these test results and the log-cumulative hazard plots. If there is a significant amount of censoring, then external data, clinical plausibility and expert judgement should be used to assess the suitability and external validity of the alternative models. If the analysis completed for Step 2 suggests that the proportional hazards assumption is reasonable and a PH modelling technique is used, the HR estimate should be taken from the relevant parametric model fitted with treatment group as a covariate, and different scenarios should be considered regarding the treatment effect over the extrapolated period.

Step 4. Based on the above analysis the most appropriate survival models should be selected for the base case analysis. The assessment of appropriateness should take account of the fit of the models to the observed data, and the plausibility of the extrapolated portion of the models. Similar types of models (with 'type' defined as the same parametric distribution) should be used for the different treatment arms unless there is strong evidence to suggest an alternative is more plausible. Where there is more than one plausible set of models, the alternatives should be included in the economic model as scenario sensitivity analysis, and for each scenario uncertainty around the parameter estimates within the chosen models should be incorporated in probabilistic sensitivity analysis. This allows the impact of choosing different models on cost-effectiveness results to be demonstrated, which provides decision makers with more information upon which to base their recommendations

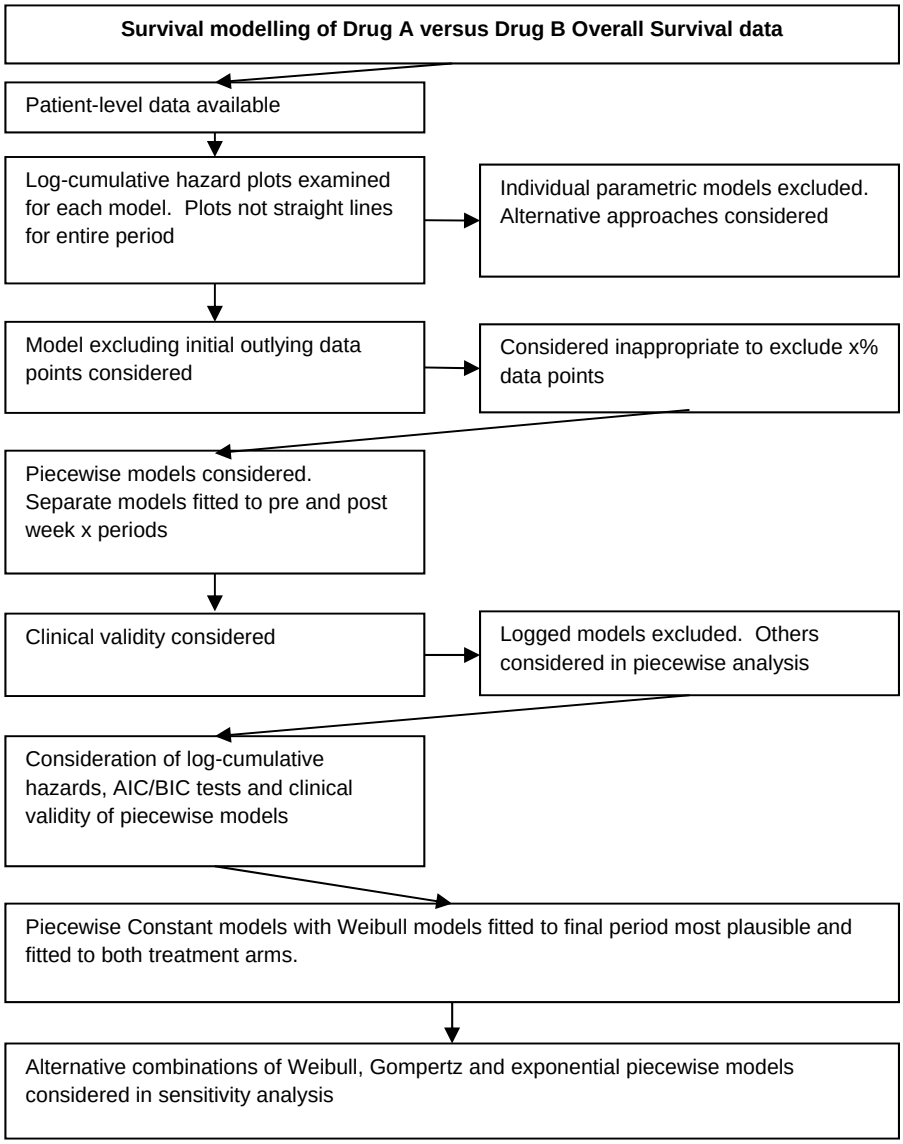
Figure 3: Survival Model Selection Process Algorithm



6.2 MODEL SELECTION PROCESS CHART

Figure 4 presents a Survival Model Selection for Economic Evaluations Process (SMEEP) Chart, that could be completed by manufacturers within their submission, to demonstrate the processes they went through when analysing their data. Figure 4 is completed for fictitious Drugs A and B for disease Y as an example, to demonstrate the type of information that should be included. Presenting such a chart as a matter of course whenever an economic evaluation incorporates survival estimates increases transparency, allows AGs to critique the survival analysis conducted by the manufacturer, and promotes consistency between Appraisals. Charts should be completed separately for progression-free survival and overall survival (or whatever time periods are being estimated).

Figure 4: Survival Model Selection For Economic Evaluations Process (SMEEP) Chart: Drug A and Drug B for Disease Y



7. REFERENCES

1. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves, BMC Med Res Methodol 2012;12:9
2. National Institute for Health and Clinical Excellence, Guide to the Methods of Technology Appraisal, June 2008, accessed from http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf, 9 March 2011.
3. Billingham, L.J., Abrams, K.R. and Jones, D.R. Methods for the analysis of quality-of-life and survival data in health technology assessment. Health Technology Assessment 1999;3;10.
4. Collett, D. Modelling survival data in medical research (2nd ed.), Boca Raton: Chapman & Hall/CRC, 2003.
5. Panageas, K.S., Ben-Porat, L., Dickler, M.N., Chapman, P.B., Schrag, D. When You Look Matters: The Effect of Assessment Schedule on Progression-Free Survival. *Journal of the National Cancer Institute* 2007; 99; 6: 428-432.
6. Breslow, N.E., Covariance analysis of censored survival data. *Biometrics* 1974; 30: 89-100.
7. Royston, P. and Parmar, M.K.B. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21:2175-2197; 2002.
8. Jackson, C.H., Sharples, L.D., Thompson, S.G. Survival models in health economic evaluations: Balancing fit and parsimony to improve prediction. *The International Journal of Biostatistics 2010*; 6;1;34.
9. Ouwens, M.J.N.M., Philips, Z., Jansen, J.P. Network meta-analysis of parametric survival curves. Research Synthesis Methods 2010; 1:258-271
10. Guyot, P., Welton, N.J., Ouwens, M.J.N.M., Ades, A.E. Survival time outcomes in randomised controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness., *Value in Health* In press.

11. Harrell, F.E. Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. Springer, New York, 2001.
12. Royston, P., Parmar, M.K.B., Altman, D.G. External validation and updating of a prognostic survival model. University College London Research Report 307, 2010. Available from: http://www.ucl.ac.uk/statistics/research/pdfs/rr307.pdf
13. Celgene Limited, Lenalidomide (Revlamid®) for multiple myeloma in people who have received at least one prior therapy, Single technology appraisal (STA) submission to the National Institute for Health and Clinical Excellence, 13 August 2008 revision, accessed from http://www.nice.org.uk/nicemedia/live/11937/42431/42431.pdf, 18 October 2010.
14. Celgene Limited, Response to Lenalidomide ERG Evaluation Report, January 2009, accessed from http://www.nice.org.uk/nicemedia/live/11937/43028/43028.pdf, 18 October 2010.
15. Hoyle, M., Rogers, G., Garside, R., Moxham, T. and Stein, K. The clinical- and cost-effectiveness of lenalidomide for multiple myeloma in people who have received at least one prior therapy: An evidence review of the submission from Celgene, Addendum to the report submitted on 1st September 2008, Peninsula Technology Assessment Group, Universities of Exeter and Plymouth, January 2009, accessed from http://www.nice.org.uk/nicemedia/live/11937/43015/43015.pdf, 18 October 2010.
16. Eli Lilly and Company Limited, Appeal by Eli Lilly and Company Limited in relation to the final appraisal determination for pemetrexed disodium for the treatment of malignant pleural mesothelioma, July 2006, available from http://www.nice.org.uk/nicemedia/live/11698/34984/34984.pdf, accessed 29 October 2010.
17. National Institute for Health and Clinical Excellence, Appraisal of pemetrexed disodium for the treatment of malignant pleural mesothelioma: Appeal panel decision, December 2006, available from http://www.nice.org.uk/nicemedia/live/11698/34985/34985.pdf, accessed 29 October 2010.
18. Moeschberger, M.L. and Klein, J.P. A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics* 1985;41;1:253-259.
19. Wilson, J., Connock, M., Song, F., Yao, G., Fry-Smith, A., Raftery, J. and Peake, D. Imatinib for the treatment of patients with unresectable and/or metastatic gastro-

intestinal stromal tumours – a systematic review and economic evaluation. Produced by West Midlands Health Technology Assessment Collaboration, University of Birmingham. Commissioned by NHS R&D HTA Programme. October 2003. Available from http://www.nice.org.uk/nicemedia/live/11547/32966/32966.pdf, accessed 2 November 2010.

20. Garside, R., Pitt, M., Anderson, R., Rogers, G., Dyer, M., Mealing, S. et al. The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high grade glioma: A systematic review and economic evaluation. Produced by Peninsula Technology Assessment Group, Universities of Exeter and Plymouth. Commissioned by NHS R&D HTA Programme. September 2005. Available from http://www.nice.org.uk/nicemedia/live/11619/34040/34040.pdf, accessed 1 November 2010.
21. National Institute for Health and Clinical Excellence, Final Appraisal Determination: Imatinib for the treatment of unresectable and/or metastatic gastro-intestinal stromal tumours. August 2004. Available from http://www.nice.org.uk/nicemedia/live/11547/32967/32967.pdf, accessed 2 November 2010.
22. Thompson Coon, J., Hoyle, M., Green, C., Liu, Z., Welch, K., Moxham, T. and Stein, K. Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: A systematic review and economic evaluation. Produced by Peninsula Technology Assessment Group, Universities of Exeter and Plymouth. Commissioned by NHS R&D HTA Programme. May 2008. Available from http://www.nice.org.uk/nicemedia/live/11817/41488/41488.pdf, accessed 29 October 2010.
23. Bond, M., Hoyle, M., Moxham, T., Napier, M. and Anderson, R. The clinical and cost-effectiveness of sunitinib for the treatment of gastrointestinal stromal tumours: a critique of the submission from Pfizer. Produced by Peninsula Technology Assessment Group, Universities of Exeter and Plymouth. Commissioned by NHS R&D HTA Programme. Available from http://www.nice.org.uk/nicemedia/live/12040/43430/43430.pdf, accessed 2 November 2010.
24. National Institute for Health and Clinical Excellence, Technology appraisal guidance 121: Carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma, June 2007, available from

http://www.nice.org.uk/nicemedia/live/11620/34049/34049.pdf, accessed 29 October 2010.

25. Dalziel, K., Round, A., Stein, K., Garside, R. and Price, A. The effectiveness and cost-effectiveness of imatinib for first line treatment of chronic myeloid leukaemia in chronic phase. Produced by Peninsula Technology Assessment Group, University of Exeter and Wessex Institute for Health Research and Development, University of Southampton. Commissioned by NHS R&D HTA programme. March 2003. Available from http://www.nice.org.uk/nicemedia/live/11515/32751/32751.pdf, accessed 29 October 2010.
26. Main, C., Ginnelly, L., Griffin, S., Norman, G., Barbieri, M., Mather, L. et al. Topetecan, pegylated liposomal doxorubicin hydrochloride and paclitaxel for second-line or subsequent treatment of advanced ovarian cancer. Produced by Centre for Reviews and Dissemination, University of York. Commissioned by NHS R&D HTA programme. September 2004. Available from http://www.nice.org.uk/nicemedia/live/11553/33023/33023.pdf, accessed 29 October 2010-10-29.
27. Hind, D., Tappenden, P., Tumur, I., Eggington, S., Sutcliffe, P. and Ryan, A. Technology assessment report commissioned by the HTA Programme on behalf of the National Institute for Clinical Excellence: The use of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer: systematic review and economic evaluation (review of Guidance No. 33), Addendum: Economic evaluation of irinotecan and oxaliplatin for the treatment of advanced colorectal cancer. Produced by The School of Health and Related Research, University of Sheffield. January 2005. Available from http://www.nice.org.uk/nicemedia/live/11561/33130/33130.pdf, accessed 29 October 2010.
28. Roche Products Limited, Rituximab for the treatment of relapsed follicular lymphoma, Roche submission to the National Institute for Health and Clinical Excellence, June 2007, Available from http://www.nice.org.uk/nicemedia/live/11730/38897/38897.pdf, accessed 29 October 2010.
29. Roche Products Limited, Rituximab for the 1st line treatment of chronic lymphocytic leukaemia, Roche submission to the National Institute for Health and Clinical Excellence, November 2008, Available from http://www.nice.org.uk/nicemedia/live/12039/43581/43581.pdf, accessed 29 October 2010.

30. Pfizer Limited, Single Technology Appraisal of Sunitinib for the treatment of gastrointestinal stromal tumours, October 2008, available from http://www.nice.org.uk/nicemedia/live/12040/43440/43440.pdf, accessed 29 October 2010.
31. Knight, C., Hind, D., Brewer, N. and Abbott, V. Rituximab for aggressive non Hodgkin's lymphoma: systematic review. Produced by the School of Health and Related Research, University of Sheffield, commissioned by NHS R&D HTA Programme. May 2003. Available from http://www.nice.org.uk/nicemedia/live/11505/32676/32676.pdf, accessed 1 November 2010.
32. Bagust, A., Boland, A., Dickson, R., Chu, P., Hockenhull, J. and Davis, H. Rituximab for the treatment of relapsed or refractory stage III or IV follicular non-hodgkin's lymphoma: ERG Report. Produced by Liverpool Reviews and Implementation Group, University of Liverpool. Commissioned by NHS R&D HTA programme. August 2007. Available from http://www.nice.org.uk/nicemedia/live/11730/38878/38878.pdf, accessed 29 October 2010.
33. Dundar, Y., McLeod, C., Boland, A., Walley, T., Hounsome, J., Bagust, A. et al. Rituximab for the first line treatment of stage III-IV follicular non-Hodgkin's lymphoma, produced by Liverpool Reviews and Implementation Group, University of Liverpool, commissioned by NHS R&D HTA Programme, April 2006. Available from http://www.nice.org.uk/nicemedia/live/11591/46698/46698.pdf, accessed 1 November 2010.
34. National Institute for Health and Clinical Excellence, Technology appraisal guidance 110: Rituximab for the treatment of follicular lymphoma, September 2006, available from http://www.nice.org.uk/nicemedia/live/11592/33547/33547.pdf, accessed 1 November 2010.
35. Ortho Biotec Limited, Velcade ® (Bortezomib) for the treatment of multiple myeloma patients at first relapse, manufacturer submission, July 2006. Available from http://www.nice.org.uk/nicemedia/live/11713/35151/35151.pdf, accessed 1 November 2010.
36. Green, C., Bryant, J., Takeda, A., Cooper, K., Clegg, A., Smith, A. and Stephens, M. Bortezomib for the treatment of multiple myeloma patients. Produced by Southampton Health Technology Assessments Centre, University of Southampton. Commissioned by NHS R&D HTA Programme. April 2006. Available from

http://www.nice.org.uk/nicemedia/live/11713/35150/35150.pdf, accessed 1 November 2010.

37. Liverpool Reviews and Implementation Group, Pemetrexed disodium for the treatment of malignant pleural mesothelioma: A systematic review and economic evaluation, University of Liverpool addendum to the assessment report, March 2006, Available from http://www.nice.org.uk/nicemedia/live/11698/36840/36840.pdf, accessed 1 November 2010.
38. Liverpool Reviews and Implementation Group, EGR Addendum: Pemetrexed for the first-line treatment of locally advanced or metastatic non-small cell lung cancer (NSCLC), University of Liverpool, June 2009. Available from http://www.nice.org.uk/nicemedia/live/12045/45084/45084.pdf, accessed 1 November 2010.
39. Merck Pharmaceuticals, Comments on the Assessment Report: Health Technology Appraisal, Bevacizumab and cetuximab for the treatment of metastatic colorectal cancer, April 2006. Available from http://www.nice.org.uk/nicemedia/live/11611/33882/33882.pdf, accessed 1 November 2010.
40. Tappenden, P., Jones, R., Paisley, S. and Carrol, C. The use of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer, Produced by the School of Health and Related Research, University of Sheffield, commissioned by NHS R&D Programme. February 2006. Available from http://www.nice.org.uk/nicemedia/live/11611/33924/33924.pdf, accessed 1 November 2010.
41. Gelber, R.D., Goldhirsch, A., Cole, B.F. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. International Breast Cancer Study Group. *Control Clin Trials* 1993;14:485–499.
42. Merck Pharmaceuticals. Single Technology Appraisal Submission: Erbitux® (cetuximab) for the treatment of locally advanced squamous cell carcinoma of the head and neck, Technical Appendix 2. August 2006. Available from http://www.nice.org.uk/nicemedia/live/11697/36794/36794.pdf, accessed 1 November 2010.
43. Merck Serono. Cetuximab for the treatment of metastatic and/or recurrent squamous cell carcinoma of the head and neck: Merck Serono Response to NICE Clarification Letter. October 2008. Available from

http://www.nice.org.uk/nicemedia/live/11987/42927/42927.pdf, accessed 1 November 2010.

44. Loveman, E., Jones, J., Hartwell, D., Bird, A., Harris, P., Welch, K. and Clegg, A. The clinical and cost effectiveness of topotecan for small cell lung cancer: a systematic review and economic evaluation. Produced by Southampton Health Technology Assessments Centre, University of Southampton. Commissioned by NIHR HTA Programme. March 2009. Available from http://www.nice.org.uk/nicemedia/live/12021/44800/44800.pdf, accessed 1 November 2010.