Survival Analysis and Extrapolation Modeling of Time-to-Event Clinical Trial Data for Economic Evaluation: An Alternative Approach
Adrian Bagust and Sophie Beale

Med Decis Making published online 30 July 2013
DOI: 10.1177/0272989X13497998

The online version of this article can be found at:
http://mdm.sagepub.com/content/early/2013/07/30/0272989X13497998

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for Medical Decision Making

Additional services and information for Medical Decision Making can be found at:

Email Alerts: http://mdm.sagepub.com/cgi/alerts
Subscriptions: http://mdm.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - Jul 30, 2013

What is This?
Survival Analysis and Extrapolation Modeling of Time-to-Event Clinical Trial Data for Economic Evaluation: An Alternative Approach

Adrian Bagust, BSc, Sophie Beale, BSc, MSc

A recent publication includes a review of survival extrapolation methods used in technology appraisals of treatments for advanced cancers. The author of the article also noted shortcomings and inconsistencies in the analytical methods used in appraisals. He then proposed a survival model selection process algorithm to guide modelers’ choice of projective models for use in future appraisals. This article examines the proposed algorithm and highlights various shortcomings that involve questionable assumptions, including researchers’ access to patient-level data, the relevance of proportional hazards modeling, and the appropriateness of standard probability functions for characterizing risk, which may mislead practitioners into employing biased structures for projecting limited data in decision models. An alternative paradigm is outlined. This paradigm is based on the primacy of the experimental data and adherence to the scientific method through hypothesis formulation and validation. Drawing on extensive experience of survival modeling and extrapolation in the United Kingdom, practical advice is presented on issues of importance when using data from clinical trials terminated without complete follow-up as a basis for survival extrapolation. **Key words:** survival analysis; decision analysis; technology assessment.

An article by Latimer recently published in *Medical Decision Making* describes a review of the survival extrapolation methods reported in all National Institute for Health and Care Excellence (NICE) technology appraisals (TAs) that have dealt with advanced and/or metastatic cancer and were completed by December 2009. Having noted various shortcomings and widespread inconsistency in the methods that had been used, Latimer presents a survival model selection process algorithm. This algorithm is intended to encourage greater conformity in projective model development for future NICE appraisals and, by implication, in the wider domain of health technology assessment (HTA). Unfortunately, Latimer’s review and resulting algorithm do not take sufficient account of the contextual constraints that apply to the various participants in the NICE appraisal process and to HTA researchers more generally. In particular, Latimer’s algorithm involves an assumption that the modeler has full access to the individual patient data from the key trial(s). Sadly, although this is generally true for trial sponsors, these data are rarely, if ever, available to other HTA researchers who must, therefore, routinely depend on the limited information and summary charts provided in brief published reports. In response, we identify several practical shortcomings in the Latimer algorithm and outline an alternative framework that has been developed from a perspective that owes more to empirical science than to statistical theory.
BACKGROUND

The NICE technology appraisal program issued its first determination in 2000 and, by January 2013, had published 273 separate appraisals. The TA process has developed over time and now comprises 2 separate elements: the single technology appraisal (STA), involving a single novel treatment or indication compared with current best practice, and the more complex multiple technology appraisal (MTA), where several similar treatments for the same indication in the same patient population are compared. In each case, the sponsor or manufacturer of the technology is invited to submit evidence of clinical effectiveness and cost-effectiveness, which is critiqued by an independent academic group. Frequently, the main evidence of clinical benefit is derived from 1 or 2 randomized controlled trials (RCTs) that have been undertaken by the manufacturer. Invariably, these trials have limited follow-up, and some form of extrapolation beyond the available data is required to facilitate a meaningful assessment of cost-effectiveness.

As part of their written evidence submission, manufacturers carry out data analysis to identify a preferred formulation for survival projection and report their findings. However, they do not normally make the patient-level trial data available to independent academic groups, including those researchers contracted to review the evidence submitted to NICE in support of applications for product approval. Therefore, although an algorithm may have some limited merit in directing manufacturers to standardize their own modeling methods, an alternative pragmatic approach is required to assist researchers who do not have the benefit of access to the primary source of data.

ALGORITHM CRITIQUE

The survival model selection process algorithm proposed by Latimer\(^1\) suffers from several important deficiencies that call into question its appropriateness in the NICE TA context as well as for the wider HTA community.

Patient-Level Data

The algorithm proposed by Latimer\(^1\) begins with the assumption that patient-level data are available for analysis before proceeding to suggest development pathways to identify and calibrate a suitable projection model. However, as noted previously, manufacturers only rarely make patient-level data available to independent academic groups, including those researchers contracted to review the evidence submitted to NICE in support of applications for product approval. Therefore, although an algorithm may have some limited merit in directing manufacturers to standardize their own modeling methods, an alternative pragmatic approach is required to assist researchers who do not have the benefit of access to the primary source of data.

Log-Cumulative Hazard Plots

Latimer\(^1\) recommends the use of log-cumulative hazard plots v. log of time as the primary means of deciding between different approaches to model
selection. This follows a traditional approach, aimed primarily at confirming proportionality of hazards between treatment arms (discussed by Bellera and others\(^3\)). However, in practice, this method can be unhelpful since it tends to distort Kaplan-Meier data by compressing long-term differences but exaggerating short-term effects so that visual comparison is difficult, subjective, and potentially misleading. In addition, it only yields a linear plot for constant risk data (from an exponential distribution) in the situation where constant risk applies over the whole duration of the trial. In clinical trials, this is unusual, as the design of most trials naturally generates varying hazards as a new treatment initially accrues additional outcome benefit, before reaching a maximum effect, and then subsequently losing effect in the long term. Figure 1 illustrates a comparison between 3 response patterns to short-term chemotherapy in which the cumulative hazard functions initially diverge before following the same constant long-term hazard. The cumulative hazard plot reveals the consistency of long-term effect, whereas the log-log plot exhibits different nonlinear trajectories without any obvious relationship. In general, it is more informative to examine the cumulative hazard plot against time since this allows easy recognition of linear trends that may be present over any time period and also clearly indicates, by nonlinear trends, periods when the hazard may be increasing or decreasing. Thus, a cumulative hazard plot will distinguish between exponential and Weibull trends, which both appear as straight lines in a log-log plot.

Proportional Hazards Models

Latimer’s algorithm\(^1\) supports the use of proportional hazard models for extrapolation in cases where log-log plots of survival are parallel. There are 2 possible reasons to pursue such an option: first, to improve the fit of the model by taking account of relevant patient covariates and, second, to develop a single parametric model that distinguishes between 2 trial arms only by use of a binary “switch” variable to adjust one of the model parameters. In practice, a researcher who does not have access to the individual patient-level data set, including all relevant baseline characteristics and survival information, would not be able to carry out a reliable covariate analysis. It is generally unwise to apply a joint model to clinical trial data for 2 treatment arms since the design of most trials provides a comparison of treatments with different mechanisms of actions and, as a consequence, different patterns of event hazard over time, so it is unlikely that a single functional form will match both trial arms successfully. The effect of employing such models is generally to bias the estimated survival in both arms as the parameter estimates are necessarily compromised away from the best fit for either arm. The presumption should be against joint modeling of treatment arms unless modeling the trial arms independently reveals that functional forms and parameter estimates are closely aligned. Nonetheless, the appropriateness of each separate functional form needs careful justification, from both the available data and other sources (such as clinical experience and, if available, patient registries).
Complete v. Incomplete Data

Latimer’s model selection algorithm\(^1\) provides 2 sets of criteria for guiding the selection of a specific model formulation, one for complete and the other for incomplete survival data. This appears to be anomalous since the only meaningful motivation for projective modeling is that the trial data are incomplete. If the data in either arm of the trial are complete (i.e., all patients have been followed up until the study event has occurred or the patient has withdrawn from the trial), then no additional explanatory power is generated by fitting a model to the data; the Kaplan-Meier estimates of survival should be used directly, since fitting a model adds only additional potentially arbitrary assumptions and, therefore, additional uncertainty, without any material gain.

Use of AIC/BIC

It is important to recognize that the use of information criteria statistics (Akaike information criterion [AIC] and Bayesian information criterion [BIC]) is of questionable value in assisting in the selection of the most appropriate functional form for a projective model. These measures provide a numerical measure of the match between observed data and model estimates across the available trial follow-up but give no indication of the relative merits of competing models when used for extrapolation. The calculated criteria values have no absolute interpretation, and differences between values only serve to rank competing model formulations (without indicating whether the magnitude of differences is important) on the optimistic assumption that the “correct” model is among those being considered. Comparison of traditional model residuals frequently reveals systematic bias in all “standard fitted models” (exponential, Weibull, Gompertz, gamma, log-logistic, lognormal), indicating that none are in fact appropriate representations of the available data.

Standard Models

Latimer\(^1\) recommends that all “standard models” should be considered and demonstrated to be unsuitable, before any alternatives are employed. Although this is common practice (and convenient when using commercial statistical software), it requires no consideration of the assumptions implicit in such analyses or whether “standard models” are appropriate for projecting beyond the available data. Curve fitting and the statistical tests used to assess relative “goodness of fit” between candidate formulations are essentially descriptive, in that they relate solely to the extent of correspondence between the available data and the calibrated standard function. It cannot be presumed that such a mechanistic process will yield clinically or physiologically credible results when projected into the future. In addition, the common practice of fitting curves to the whole of the available data may not be sensible, especially for analysis of clinical trials where phased alterations to patient treatment are an intrinsic part of the trial design. This inevitably conflicts with standard mathematical formulations, which are essentially smooth continuous functions, without abrupt alterations in trends.

In the real world, several aspects of RCTs may contribute to the difficulties that prevent a good “model fit.” These include the following:

a) The action of a newly prescribed drug may take some time to achieve its full effect, partly due to the pharmacokinetic/dynamic profile of the drug and partly due to the time required for the active agent to achieve its full effect at the target site(s). Conversely, when the period of active treatment comes to an end, its effects may only dissipate gradually over several weeks. This is also relevant where there is no washout period between completing a prior course of treatment (such as following first-line chemotherapy) and commencing the trial intervention.

b) Additional confounding may be introduced by the availability of subsequent courses of active treatment following discontinuation of the study treatment. This further complicates the dynamic nature of the event hazard rate.

c) There is also the possibility that the patient population is essentially heterogeneous in relation to the event risk of interest, leading to progressive survivor bias as members of one subgroup die at a faster rate than other patients.

As a consequence, it is not surprising that fitting a standard parametric survival function to the full clinical trial data set rarely produces a satisfactory correspondence to the observed survival trajectory.

The computational convenience of most “standard functions” is not accompanied by obvious biomedicai causal models, which might lend external credibility to their use and provide insight into the modes of action of interventions at the individual or population levels. Therefore, there is no objective justification for according primacy to a small number of theoretical distributions over a dispassionate
examination of the available data from properly conducted clinical trials.

Plausibility

Latimer prefers use of models that “provide a plausible extrapolation... based on external data and clinical expert opinion” and are “based on internal and external plausibility.” In practice, this advice is difficult to interpret; internal plausibility may be no more than the relative measures of “good fit” discussed above, and both external data and opinion sources are vulnerable to selective citation. The concept of plausibility is clearly important in model selection but requires careful definition, recognizing that the relevant concerns may differ in each specific context.

PRACTICAL ADVICE FOR MODEL DEVELOPMENT

In this section, several suggestions and cautions are offered to aid the analyst. This advice is based on experience gained from undertaking multiple health technology assessment projects that have required projective modeling.

Examine the Cumulative Hazard Plot

Convert the event time data of the Kaplan-Meier (K-M) survival analysis (i.e., the survival value and time at the bottom of each downward step in the plot) to cumulative hazard data, using the following conversion: Cumulative hazard = – log (estimated K-M survival).

Plot the cumulative hazard data against time in trials, for both (or all) trial arms together, and consider whether there is evidence of long-term linear trends in each arm in the latter part of the data, when transient effects have dissipated. Our experience and the principle of parsimony suggest that the exponential distribution should be considered the default parametric function for long-term survival projection and that clear contrary evidence should be required before other options are considered. Circumstances where deviations from the exponential distribution may be more likely include trials with a very long follow-up period, during which accelerating background mortality risk with age can be expected, or if there is a realistic expectation of genuine cure of the primary disease. If a long-term constant hazard trend is indicated, then an exponential model may be fitted to the remaining data after left truncation of the data set at the point when the trend is clearly established.

Parallel Cumulative Hazard Trends

A special case arises when long-term trends in each arm of a clinical trial appear to exhibit very similar hazard rates. This may imply that, after the intervention treatment is completed, withdrawn, or ceases to deliver additional benefit, an identical long-term risk trajectory applies regardless of treatment. If this can be established, then the difference in mean survival can be accurately estimated by comparing the area under the survival plot for each trial arm, up to the point at which survival risk becomes identical. However, the full equivalence of long-term trends can only be confirmed if it is possible to obtain a Kaplan-Meier analysis of the truncated data (i.e., the period during which similar hazard rates appear in trial arms). This is especially important in cancer RCTs, in which only preprogression and overall survival are normally reported; direct analysis of postprogression survival may resolve the issue (e.g., Figure 2,\(^4\) in which no difference between trial arms is discernible), although it should be noted that some minor secondary difference in postprogression survival can still arise if there is a difference between progression fatality rates in the trial arms.

Figure 3 illustrates the value of the cumulative hazard plot for assessing the duration of benefit of a new treatment. In this case, the hazard in the...
comparator appears to be fixed in all time periods (i.e., exponential trend), whereas the new treatment suppresses mortality for about 3 months, after which a common long-term hazard becomes evident. In this instance, a good estimate of the expected health gain can be obtained without any resort to projective modeling.

Clinical Protocol Effects

Reported evidence from a clinical trial, usually in the form of a Kaplan-Meier plot, is frequently influenced by aspects of the trial design that should be confirmed, where possible, by careful examination of the trial protocol. The method of selecting patients for the trial may be influential in the first few weeks of the trial, so that exclusion criteria may ensure that the risk of any patient experiencing a target event (death, disease progression, or acute crisis) is artificially suppressed for several months. Conversely, trials that require included patients to have had an acute event in the preceding 24 hours often show elevated fatalities in the first few days before stability is fully restored. Such effects occurring in either or both trial arms (near-zero risk initially or very high early attrition) may not reflect long-term survival and should be carefully investigated before fitting any parametric function.

Another common protocol effect, especially in cancer trials, concerns the timing of periodic disease evaluations, which generate regular “steps” in the Kaplan-Meier plot of confirmed disease progression, corresponding to the timing of the protocol-driven tests and clinical assessments. The underlying risk profile is therefore represented only by the data points at the end of each disease assessment phase, since it is only possible to assess accurately how many patients remain progression free under the same reporting conditions (i.e., combining both mandated and opportunistic detection) at these time points. Improved estimation may be achieved by modifying the trial data by, for example, use of interval censoring or by modeling the mandated and opportunistic components of disease progression separately. However, in most advanced cancer trials, the data for progression-free survival is quite mature, so that the scope for extrapolation bias is much less than that for overall survival.

The use of compound outcome measures is potentially problematic and is common in trials relating to particular diseases such as cancers (progression-free survival ends at death or confirmed disease progression) and cardiovascular disease (measured by combinations of death, acute myocardial infarction, cerebral hemorrhage or infarction, etc.). These require careful consideration; when separate components with very different incident risks and temporal trajectories are combined into a single measure, the resulting survival plot is unlikely to conform to any standard statistical function and, in the long term, is likely to revert to a single trend governed by the dominant component of the compound outcome measure. However, in advanced and metastatic cancer, it is generally the case that most deaths are attributable to the cancer itself or to adverse effects of anticancer therapies, so that overall survival is the most reliable and objective outcome measure routinely available.

The censoring rule applied in RCT survival analysis is generally overlooked when projecting survival. Most pivotal trials providing evidence for drug approval or reimbursement are terminated as soon as a significant treatment effect is detected, resulting in right-censoring of incomplete patient records at a particular date. It is conventional in most published results to consider the operative timing of such censoring to be at the time of the last patient observation or the time of any subsequent event. This is justified on the grounds that this approach provides conservative estimates of the median survival. However, this is inappropriate for Kaplan-Meier analysis since it violates the fundamental assumption of uninformative censoring by systematically underestimating
the time when censored patients are at risk of death (or other event), while retaining any events that have occurred since the last patient observation. As a consequence, final period hazards are overestimated and trend lines distorted. In trials that are terminated at a fixed date, this is often visually evident as sudden downward movements in the Kaplan-Meier plot at the end of the observed data (as demonstrated by the simulation results presented in Figure 4), corresponding to an equivalent unexpected upward movement in the cumulative hazard trend. Where this is observed, a reanalysis of the data using the correct censoring times should be requested from the data custodian. If this is not possible, consideration should be given to reestimating long-term trends, excluding the implicated final data points, at least as a sensitivity analysis.

Possible Subgroups

Widespread use is made of 2- and 3-parameter “standard functions” for projective models without consideration of any credible mechanism by which these forms may be generated from population characteristics or disease processes. Furthermore, there is often no attempt to match biological causes by analogy to time-to-event processes modeled in other disciplines, such as engineering, operational research, or information science. There are, however, some insights available with respect to the exponential and Weibull functions. If a population is heterogeneous for survival, it is inevitable that over time, those patients at higher risk will tend to die earlier so that those with a better prognosis will come to predominate in the surviving group. This means that regardless of the nature of the risk profiles of the population subgroups, the combined hazard function will be monotonically decreasing. This corresponds with the observation that long-term trends frequently exhibit Weibull-like characteristics with a shape parameter less than 1. In 1982, Jewell considered the characteristics of mixtures of exponential distributions and proved that “any Weibull distribution with shape parameter less than 1 arises as a mixture of exponentials. Also the exponential distribution itself arises as a mixture of Weibull distributions with fixed shape parameter p, as long as p>1.” It is reasonable, therefore, to consider that where a Weibull function appears to provide a good fit to long-term clinical trial data, this may arise as a result of differential survival among 2 or more subpopulations subject to fixed long-term hazards.

An example of such a situation occurred in the appraisal of a novel treatment for metastatic malignant melanoma. Despite a very high early mortality rate among trial patients, it was evident that some patients appeared to survive much longer than would normally be expected in late-stage cancers. American Joint Committee on Cancer (AJCC) registry data for advanced malignant melanoma displayed a similar phenomenon, and clinical advisers remarked that they were aware that a small group of patients do indeed survive for uncharacteristically long periods. Exploratory modeling showed that a mixture of 2 exponential distributions (19% with mean survival
of 11 years and 81% with mean survival of 11 months) very closely matched the AJCC advanced malignant melanoma data as shown in Figure 5. It appears that similar exploratory modeling may be useful in lending credibility to some projective models, as well as in furnishing new hypotheses for targeting research to identify patients most likely to benefit from treatment.

**PRAGMATIC MODELING PARADIGM**

When approaching the task of identifying a suitable method for estimating a time-to-event outcome from restricted data, several basic principles should be considered:

- Only attempt projective modeling to estimate lifetime survival if full trial data are not available.
- Concentrate attention on using the later stable portion of the available data as the basis for projecting outcomes beyond the available data, rather than simply aiming to achieve a good fit over the whole data set.
- Accept the primacy of trial data over any a priori assumptions or analytical constraints applied to simplify model fitting.
- Appreciate that trial data are subject to the limitations imposed by the study protocol, which may require use of analytic methods designed to minimize protocol-related biases.
- Recognize that, in considering the outcomes of any novel intervention, all types of treatment effect should be considered potentially legitimate.
- Prefer methods of extrapolation, which can be associated directly, or causally, with known or likely modes of action over methods based on technical convenience.

In effect, the modeler should follow a classic process of scientific evidence review leading to hypothesis formulation. Typically, the scientific method involves developing causal hypotheses to explain observed phenomena. Most clinical drug trials have limited objectives, centered on demonstrating the superiority of a new intervention over an existing form of treatment using a narrow set of outcome measures (such as measured response to treatment or median survival). However, the HTA researcher needs to answer wider questions about patient outcomes over a lifetime and therefore requires a basis for determining the most appropriate approach to employ when extrapolating trial results. A formulaic approach that selects a statistical function from a predefined limited pool of convenient forms provides no insights into the nature of the disease and treatment processes generating the observed trial results. However, following the proposed pragmatic paradigm can lead to important conclusions about the extent and timing of treatment effects, their duration/persistence, and the significant likelihood of heterogeneity of effect leading to possible subgroup effects. Candidate models based on scientific evidence review should then be tested against the source data and, where possible, other related data sources. The resulting formulation should then be subjected to external review for feasibility and clinical credibility. If found wanting, inadequate hypotheses and the extrapolation model should be revised and the review-development cycle revisited.

**CONCLUSION**

Although Latimer’s review of modeling methods reported in previous NICE appraisals is useful and interesting, his proposed algorithm to guide model selection is not helpful as it contains questionable assumptions and potentially misleading advice. We are concerned that such a schema risks perpetuating the notion that extrapolation modeling in survival analysis can be reduced to a set of simple mechanistic steps. Instead, we consider that the handling of clinical trial data should be subject to the full rigor of experimental scientific enquiry, with an openness to the possibility that analysts’ preconceptions and assumptions may be inadequate and inappropriate.

**REFERENCES**

5. NICE technology appraisal TA269, vemurafenib for treating locally advanced or metastatic BRAF V600 mutation-positive malignant melanoma. Evidence Review Group report Figure 7.