
Article:
Kearns, Benjamin, Stevenson, Matt, Triantafyllopoulos, Kostas et al. (Accepted: 2022) Dynamic and flexible survival models for extrapolation of relative survival: a case-study and simulation study. Medical Decision Making. ISSN: 1552-681X (In Press)

Dynamic and flexible survival models for extrapolation of relative survival: a case-study and simulation study.

Running title: Dynamic and flexible relative survival models

Word count (excluding title page, abstract, references, table): 4135

Structured abstract

Background

Extrapolation of survival data is a key task in health technology assessments (HTAs), which may be improved by incorporating general population mortality data via relative survival models. Dynamic survival models are a promising method for extrapolation which may be expanded to dynamic relative survival models (DRSMs), a novel development presented here. There are currently neither examples of dynamic models in HTA nor comparisons of DRSMs with other relative survival models when used for survival extrapolation

Methods

An existing appraisal, for which there had been disagreement over the approach to survival extrapolation, was chosen and the health economic model recreated. The sensitivity of estimates of cost-effectiveness to different model choices (standard survival models, DSMs and DRSMs) specifications was examined. The appraisal informed a simulation study to evaluate DRSMs with relative survival models based on both standard and spline-based (flexible) models.

Results

Dynamic models provided insight into the behaviour of the trend in the hazard function and how it may vary during the extrapolated phase. DRSMs led to extrapolations with improved plausibility for which model choice may be based on clinical input. In the simulation study, the flexible and dynamic relative survival models performed similarly and provided highly variable extrapolations.

Limitations

Further experience with these models is required to identify settings when they are most useful and the accuracy of their extrapolations.

Conclusions

Dynamic models provide a flexible and attractive method for extrapolating survival data and facilitate the use of clinical input for model choice. Flexible and dynamic relative survival models make few structural assumptions and can improve extrapolation plausibility, but further research is required into methods for reducing the variability in extrapolations.

Introduction

Health technology assessment (HTA) is the scientific evaluation of health technologies and informs decisions about if a health technology should be funded. For consistent decision making, all relevant costs and consequences associated with the appraised technology should be included in HTA. When the treatment impacts on survival it is important that lifetime outcomes are included in the assessment (1). Estimates of lifetime mean survival typically require extrapolations of incomplete survival functions. These estimates can be key drivers of estimates of cost-effectiveness, and hence funding decisions (2). This illustrates the importance of using appropriate methods for extrapolation.

A recent review of methods for extrapolating survival data in cancer appraisals concluded that current approaches were "suboptimal", with an over-reliance on common survival models, which may not adequately capture the complexities of hazard functions that are expected to arise from clinical trials (3). Dynamic survival models (DSMs) have recently been suggested as flexible models for the analysis and extrapolation of survival data (4). These may be viewed as relaxing the structural assumptions of common survival models by allowing their parameters to vary over time, with this temporal variation modelled by a time series. A particular advantage of DSMs is that extrapolations are based on all the data, whilst simultaneously giving more weight to more recent observations. This resolves the disagreement in the literature over how much evidence should be included when generating extrapolations(5-8). Despite these advantages of DSMs, there is a dearth of examples of their use in HTA.

Another approach to improve extrapolations is via the incorporation of external long-term evidence, such as general population mortality data (9-13). In particular, additive relative survival models decompose the overall hazard function into the sum of disease-specific (or 'excess') hazards and general population hazards. Extrapolations are obtained for the former, and the additive structure ensures that the overall hazard function never falls below the general population hazards. Models

for the disease-specific hazard function include standard parametric models and flexible spline-based models (14-17). In addition, DSMs may be used, providing dynamic relative survival models (DRSMs), a novel method that has not previously been evaluated.

This manuscript has two primary objectives. The first is to demonstrate the use of DSMs and DRSMs in HTA via a re-analysis of an existing NICE appraisal. For this appraisal estimates of cost-effectiveness were sensitive to the choice of extrapolating model for overall survival (OS), and a key critique of the original extrapolations was that they fell below those of the age-matched general population. The second objective is to perform a simulation study, informed by the appraisal, to compare the performance of relative survival models.

Methods

The code used for both the case-study and simulation study is available online (https://github.com/BenKearns/RelativeSurvival) and provides additional information.

Case study: squamous non-small-cell lung cancer

The existing HTA was a submission to NICE as part of their TA programme (18). A NICE committee considers both the company submission and the independent evidence review group (ERG) critique of this as part of their decision-making process. The NICE committee provides recommendations on if the technology is judged to be cost-effective and hence whether the technology should be recommended for routine use. For this appraisal, the population of interest was people with previously treated locally advanced or metastatic (stage IIIB or IV) squamous non-small-cell lung cancer. The intervention was nivolumab and the sole comparator in the company's submission was docetaxel. The main evidence source was the phase III trial CheckMate-017 (NCT01642004) which compared nivolumab (n = 135) against docetaxel (n = 137) for the population of interest (whose previous treatment was with platinum combination chemotherapy)(19). Patient follow-up was between 11 and 24 months. At the end of follow-up there had been 86 (63.7%) and 113 (82.5%) deaths in the nivolumab and docetaxel arms, respectively. The primary outcome measure was OS. Evidence on effectiveness came solely from this trial and there was no treatment switching in the data used in the company's original submission.

For both OS and progression-free survival (PFS), the company based their approach to extrapolation on the guidance in NICE TSD 14 (20). The assumption of proportional hazards was checked both visually and via significance tests. The company considered both standard survival models and Royston-Parmar models (RPMs) (21), with up to two internal knots modelled on the hazard, normal and odds scales (corresponding to extensions of the Weibull, lognormal and log-logistic models,

respectively) and Akaike's information criteria (AIC) for goodness of fit. For OS, the assumption of proportional hazards appeared to hold, with a log-logistic model used for docetaxel. The treatment effect for nivolumab was modelled as a fixed hazard ratio of 0.59. For PFS, the proportional hazards assumption was judged to be violated. Hence the company modelled both treatments using an RPM with two internal knots on the hazard scale. The probabilistic base-case incremental cost-effectiveness ratio (ICER) arising from this approach was £86,000 (all ICERs discussed in this manuscript are given to the nearest £500 and are per QALY gained), with a survival gain of 1.31 years for nivolumab (18). This value was robust to alternative approaches to extrapolation for PFS, but not for OS. For example, when varying the hazard ratio across its plausible range the ICER varied from £55,000 to £169,000.

The independent ERG were critical of the company's OS extrapolations, in particular the fact that the extrapolated hazard eventually fell below that of the age-matched general population was deemed to be "wholly implausible, and inconsistent with any clinical evidence of treating metastatic disease" (22). The ERG contended that the extrapolated hazard for OS was likely to increase over time due to ageing. Despite this, they extrapolated a constant hazard over time (using an exponential model). This was fit from 40 weeks (9.2 months) of follow-up (a temporal subset of the data), with the ERG suggesting that this cut-off was supported by the data. The ERG's approach to OS extrapolation increased the company's base-case ICER from £86,000 to £132,000, whilst the estimated lifetime survival gain more than halved, from 1.31 to 0.64 years. In response, the company amended their extrapolation approach to cap the extrapolated hazard rate so that it never fell below that of the corresponding general population. The company's revised base-case ICER was £92,000, with a survival benefit of 1.16 years (23). However, the ERG remained critical of the company's revised approach as not reflecting an anticipated long-term increase in hazards due to the effect of ageing (24).

Hence, the approach to extrapolating OS was identified as both a key area of uncertainty and a key driver of estimates of cost-effectiveness. The company fit survival models to all the available data and extrapolated a decreasing trend in the hazard. In contrast, the ERG fit a survival model to a subset of the available data and extrapolated a constant value (no trend), whilst also criticising the company's original extrapolations for eventually falling below that of the age-sex matched general population. The company in turn criticised the ERG's approach as ignoring the trend in the hazard observed in the trial and lacking robustness by not using all the available data.

Case study: re-analysis of the clinical effectiveness data

Data on OS were digitised from the pivotal trial publication (19) using Engauge digitiser (25). These digitised data were used to replicate the original individual patient data using the algorithm of Guyot and colleagues (26, 27). For consistency with the original company submission, initially both current practice and RPMs are considered for the docetaxel arm (providing the baseline hazard function), with DSMs introduced later. A fixed hazard ratio is used for the nivolumab treatment effect.

Within-sample goodness of fit is measured using AIC (there were no substantial differences when using Bayesian information criteria). Another measure, the inverse evidence ratio (IER) is also used to facilitate model comparisons. The IER is a measure of how plausible a model is, relative to the `best' model (which has the minimum information criteria). Let IC_m be the information criteria value (such as AIC) for model m, with minimum value IC^*. The IER for model m is then $\exp(-0.5 * [IC_m - IC^*])$, and will be 100% for the best fitting model, whilst values for poorly fitting models will be close to zero (28). Hence the IER provides an interpretable scale for comparing model fit. Values are shown in supplementary Tables 1 and 2 and demonstrate that the log-logistic model is the best-fitting for both the standard models and the RPMs. Estimates of the hazard function from the

second-best fitting RPM (four internal knots, odds scale, results not shown) were visually very similar to the log-logistic model for both the within-sample and extrapolated periods.

Two DSMs are evaluated: a local trend and a damped trend model (see the supplementary material for model specification). Both may be viewed as modelling the log-hazard as a linear function of log-time. They differ with regards to the behaviour of their extrapolations; a local trend model extrapolates the trend in the log-hazard indefinitely whilst for the damped trend model the extrapolated trend decreases as the extrapolation time horizon increases. Three DRSMs were evaluated: local trend, damped trend, and local level implementations (see the supplementary material for descriptions). As DRSMs formally incorporate external evidence on general population mortality they are anticipated to provide more plausible extrapolations than DSMs for this case-study.

To perform cost-effectiveness analyses, the company's three-state partitioned survival analysis economic model was replicated in R, assuming a (lifetime) 20-year time horizon with a 1-week time cycle. Utility data and resource use were primarily taken from CheckMate-017 (18). The two alive health states of 'stable' and 'progressed' disease were assigned utilities of 0.750 and 0.592 (with standard deviations of 0.236 and 0.315), respectively. Everybody started in the stable health state. Results are based on a probabilistic sensitivity analysis with 2,000 iterations to account for non-linearities in the model inputs. The model structure and inputs matched those reported in the original appraisal (18). Further details on the health economic model are provided in the supplementary material.

Simulation study

An additive hazards relative survival log-logistic model was used as the data-generating mechanism for the simulation study. To ensure that this mechanism was clinically plausible, it was obtained by

fitting a log-logistic model to the case-study data (docetaxel arm), simulating from this model, and incorporating the (age-matched) general population hazard. For each individual, three times were simulated: a survival time from the log-logistic model; a survival time from the general population hazards (assuming a uniform distribution of deaths within a year); and a censoring time uniformly distributed between five and six years. This length of follow-up was chosen to ensure that there was sufficient data that included the turning-point in the hazard function. The observed survival time was set to the minimum of the three sampled times (with event status similarly set). For this study, 200 simulations were performed, with each having a sample size of 300. Estimates of the 'true' hazard function were based on the mean of 10 million simulations. Five models were considered: a log-logistic relative survival model, DRSMs with either a local or damped trend, and two flexible relative survival models. These both use cubic splines to model the excess hazard and vary with how the model is specified. One uses the specification introduced by Nelson and colleagues (hereafter 'Nelson relative survival' [NRS]), the other may also be written as a flexible mixture cure (FMC) model; for both models further details are provided by Jakobsen and colleagues (15).

The estimand was the mean of the natural logarithm of the time-varying hazard function. The primary performance measure used was the mean (of the) squared error (MSE), with bias as a secondary performance measure. For MSE smaller values indicate better model performance, for bias this is indicated by values closer to zero. To avoid results being unduly influenced by implausibly large extrapolations, hazard estimates were capped to not exceed one. Bias may be viewed as estimating how close to the truth estimates are on average, whilst MSE measures both bias and variability in estimates. Further details on the performance measures are available in the supplementary material.

Results

Case-study

Estimates of the trend in the hazard function over time, along with the uncertainty in these estimates are shown in Figure 1 for the two DSMs. This is of particular importance as there was disagreement over the assumed trend at the end of follow-up, with the company modelling a decreasing trend and the ERG modelling no trend. The trend estimate from both DSMs is initially positive followed by a decrease. For both models the trend becomes negative at about half a year. For the local trend model the trend estimates continue to decrease, albeit with a large degree of uncertainty. For the damped trend model the trend is almost zero after half a year, suggesting that after this time the assumption of a constant hazard may be appropriate. Figure 1 suggests that models which assume monotonicity (such as the Weibull and Gompertz) are inappropriate. In contrast, use of a log-logistic or lognormal model may be acceptable, as the hazards from these can increase then decrease. Further, the confidence intervals from both models include zero at all time points, indicating that a constant hazard model cannot be ruled out.

A visual comparison of the fit from the two DSMs along with the original company approach (log-logistic) and ERG approach (hybrid exponential) is provided in Figure 2. The observed hazard is generally unimodal, albeit with large variability due to small patient numbers towards the end of follow-up. For extrapolations, estimates of the annual hazard of all-cause mortality for the age-matched general population are also included based on 2016 UK data from the Human Mortality Database (29), assuming a starting age of 63 (the median age of participants in CheckMate 017). For the first year of follow-up estimates of the hazard function from the log-logistic and two dynamic models are visually similar, albeit the peak in the hazard is more pronounced for the log-logistic. At one year of follow-up there are only 30 people still at risk (22% of the starting sample); this small sample size may be driving the differences in model estimates after one-year. These differences continue into the extrapolated phase, with the largest decreases in the hazard function observed for

the log-logistic model. In contrast, the damped trend model extrapolates almost constant hazards; in the short-term these estimates are very similar to those from the ERG approach, but they become increasingly smaller than the ERG extrapolations as the time horizon increases. Extrapolations from the local trend model lie between the log-logistic and damped trend models, eventually falling below age-matched general population estimates at approximately 15 years; hence potentially lacking face validity.

Estimates from DRSMs are shown in Figure 3, along with the log-logistic model and ERG approach for comparison. Visually, the local level DRSM provides similar within-sample estimates to an exponential model and does not fit the data as well as the other models. Extrapolations from the local level and damped trend DRSMs are very similar to each other, illustrating that (as with the damped trend DSM) there is a pronounced dampening of the trend before the end of follow-up. After 20 years, hazards from all the DRSMs are greater than the general population estimates, implying that there is a non-negligible extrapolated excess hazard. After about ten years the local trend DRSM extrapolates an increasing hazard, suggesting that after this point the influence of ageing on the hazard function outweighs the extrapolated decrease in the excess hazard.

Table S4 compares the replication with the original company submission (using their approach to extrapolation) with the replicated model. Given that the individual patient-level data were recreated, there is in general close agreement, albeit with some under-estimation of absolute costs. This is expected, as it was not possible to include a drug acquisition cost for the progressed disease health state. Cost-effectiveness results from the dynamic models are provided in Table 1. For comparison, three replicated analyses are also shown:

- The company's original submission (extrapolation with a log-logistic model)
- Above, with extrapolated hazards capped by general population hazards.

- The ERG's hybrid approach (use Kaplan-Meier estimates up to 40 weeks, extrapolations based on an exponential fit to the remaining data).

As shown in Table 1 and Figures 2 and 3, extrapolations can differ between the five dynamic models, which affects the cost-effectiveness results. The smallest ICER occurs for the local trend DSM (£113,000). The largest ICERs arise from both damped trend dynamic models and the local level DRSM (£140,000 to £143,000). These three models all extrapolate a near-constant hazard. Variation in ICERs across the three DRSMs (£122,500 to £143,000) was slightly greater than variation between the ERG approach (£125,000) and the company submission with a cap (£140,000). Advantages of the DRSMs are first that model choice may be guided by clinical input into the likely behaviour of the long-term excess hazard, and secondly that external evidence is formally included as part of the model fitting procedure, instead of via a post-hoc adjustment. Collectively this allows for a stronger emphasis on understanding the likely behaviour of the long-term excess hazard function and the plausibility of different assumptions about this long-term behaviour. As noted, an advantage of dynamic models over hybrid models is the avoidance of the subjective choice of which data to use for the extrapolating model. Estimates of cost-effectiveness can be sensitive to this choice, as illustrated in supplementary Figure S1. Dynamic models also use all the data; with the ERG approach only a third of the original sample (45 people) contribute to extrapolations.

Simulation study

A graph of the true hazard function and the simulations from this is provided in the supplementary material (Figure S2), whilst a visual comparison of model estimates with the truth is given in Figure 4. The correctly specified log-logistic relative survival model has the smallest variation in extrapolations, but there is a persistent over-estimation which becomes more pronounced as the extrapolation time increases. Of the two flexible models (NRS and FCM), the NRS tends to over-

estimate the true hazard function, whilst the FCM under-estimates it. Of the two dynamic models, the damped trend model has less variability in extrapolations, due to the dampening of the trend. However, this dampening means that often the decrease in the excess hazard is not captured, leading to over-estimation. All the flexible and dynamic relative survival models produce highly variable extrapolations, especially when compared with the log-logistic relative survival model.

Summary MSE and bias values are provided in Table 2, with plots of these statistics over time provided in Figure S3. Consistent with Figure 4, the log-logistic model has the smallest variance, smallest bias, and lowest MSE values of all the relative survival models considered. Of the incorrectly specified models, MSE values were smallest for the NRS and trend-DRSM (values of 0.086 and 0.089, respectively) and largest for the FCM (0.202). The trend-DRSM had the smallest bias (0.127), however there was a lot of uncertainty in the bias estimates, with each model's confidence interval including the bias point-estimate for every other model (including the log-logistic).

Discussion

A motivating case-study introduced DRSMs and demonstrated the usefulness of relative survival models when extrapolating survival data. This case-study informed a simulation study which was used to compare different relative survival models. Flexible and dynamic relative survival models did not perform as well as the true model but are a potentially useful approach when the true survival model is unknown. The case-study illustrated several benefits of dynamic models. This includes combining flexible fit to the observed data with explicit modelling of the long-term trend, incorporating external data to inform extrapolations, and encoding clinical views on long-term survival via model specification.

The clinical plausibility of extrapolations is very important. Additive relative survival models ensure that the extrapolated hazard function does not fall below that of the general population. This is not the only measure of extrapolation plausibility, but it is an important one that should be considered. Different model specifications are possible for DRSMs, reflecting different assumptions about the long-term behaviour of the excess hazard. The options included here were that the excess hazard was constant, the observed trend continued until the excess hazard became zero, or the observed trend continued in the short-term, with long-term constant values of the excess hazard. This flexibility in model specification and the direct interpretation of the extrapolations is a significant advantage of DRSMs when compared with other survival models and allows for the natural inclusion of clinical opinion about both the natural history of the disease and the likely mechanism of action of treatments. Basing model choice on clinical input into the natural history of the disease is of particular use, as good within-sample goodness of fit is not a predictor of good extrapolation performance (30). A further advantage of DRSMs is that it is straight-forward to extend these to incorporate time-varying treatment effects which act on the disease-specific (excess) hazard function (see supplementary material for specification). As the focus of the manuscript was on different relative survival models, this extension was not pursued further, but it is noted that

modelling the treatment effect as applying to the overall hazard can lead to biased results as it includes the unrealistic assumption that treatment will reduce mortality that is unrelated to the disease(31).

Recreated patient-level data was used in the case-study. The recreated company submission showed close agreement with the original submission, demonstrating the usefulness of using recreated data. One limitation was that it was not possible to explore the effects of covariates on survival. In particular, when estimating relative survival it has been demonstrated that including age can lead to increased accuracy (32).

The case-study results from the DRSMs suggest that the ICER arising from the company's original approach (£88,000) is likely to be too low; depending on the long-term prognosis of patients the ICER is likely to be between £122,000 and £143,000. This range of ICERs is above the acceptable threshold for end-of life-treatment, which is typically assumed to be £50,000. Following their original submission, the company offered a discount to the cost of their treatment to lower the ICER (and so improve the possibility of a positive recommendation). The magnitude of discount required to make the treatment cost-effective will be strongly affected by the extrapolation approach used. Of the approaches evaluated here, it is not possible to definitively state which would be the preferred base-case analysis, but the use of a dynamic model which incorporates external evidence appears to be the most useful. Future research could identify the situations when the different DRSM specifications (including the modelling of the treatment effect) are the most appropriate. Relative survival models which do not bound the overall hazard by the general population hazard are also possible (16). There is uncertainty about if long-term extrapolated hazards should be bounded by the general population hazards (that is, if long-term survivors have a better prognosis than the general population) – long-term follow-up from trials would be able to provide insight into this.

The correctly specified log-logistic relative survival model provided the best extrapolations of the models considered. Alternative standard parametric relative survival models were not considered, as these typically have strong parametric assumptions. For example, the Weibull model assumes that the excess hazard is monotonic which is known to be inadequate for the simulation study. In practice, the suitability of a model with monotonic hazards may be unknown; similar work on model choice for cure models has shown that for standard parametric models, extrapolations can be sensitive to model misspecification (11). The alternative relative survival models considered in the simulation study have very weak structural assumptions, and so model misspecification is less of an issue. However, these models can provide highly variable extrapolations. Further research into reducing the variability of these extrapolations will be very useful. This could involve use of other types of external evidence, such as registry data or previous trials for the disease of interest (33). Further research could also identify if there are certain situations when one or more of the models considered are of particular benefit.

In conclusion, survival data describe the occurrence of deaths over time and so form a natural time series. This motivates the use of dynamic models, which can exploit the temporal evolution of the hazard function when generating extrapolations. These models combine flexible within-sample estimates with parsimonious models for extrapolations which have meaningful clinical interpretations. These models, along with relative survival models that incorporate external evidence on general population mortality, have potential advantages over the survival models currently used in HTA. In the simulation study of this manuscript, dynamic and flexible relative survival models had similar extrapolation performance. These models impose minimal structural assumptions and can provide good within-sample estimates. Further experience of these models is required to provide more specific guidance about the role of both dynamic models and relative survival models in HTA.

Acknowledgements

[See title page for acknowledgements.]

References

1. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013 2013 [updated April 2013; cited 2017 10/03/2017]. Available from: https://www.nice.org.uk/process/pmg9/chapter/foreword.
2. Kearns B, Stevens J, Ren S, Brennan A. How Uncertain is the Survival Extrapolation? A Study of the Impact of Different Parametric Survival Models on Extrapolated Uncertainty About Hazard Functions, Lifetime Mean Survival and Cost Effectiveness. PharmacoEconomics. 2019;38(2):1-12.
3. Bell Gorrod H, Kearns B, Thokala P, Labeit A, Stevens J, Latimer N, et al. Plausible and consistent tails: a review of survival extrapolation methods used in technology appraisals of cancer treatments Medical decision making : an international journal of the Society for Medical Decision Making. 2019.
4. Kearns B, Stevenson M, Triantafyllopoulos K, Manca A. Generalized Linear Models for Flexible Parametric Modeling of the Hazard Function. Med Decis Mak. 2019;39(7):12.
5. Latimer NR. Response to "survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach" by Bagust and Beale. Medical decision making : an international journal of the Society for Medical Decision Making. 2014;34(3):279-82.
6. Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. Medical decision making : an international journal of the Society for Medical Decision Making. 2014;34(3):343-51.
7. Latimer NR. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Medical decision making : an international journal of the Society for Medical Decision Making. 2013;33(6):743-54.
8. Kearns B, Jones ML, Stevenson M, Littlewood C. Cabazitaxel for the second-line treatment of metastatic hormone-refractory prostate cancer: a NICE single technology appraisal. Pharmacoeconomics. 2013;31(6):479-88.
9. Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information. Medical decision making : an international journal of the Society for Medical Decision Making. 2017;37(4):353-66.
10. Jackson C, Stevens J, Ren S, Latimer N, Bojke L, Manca A, et al. Extrapolating survival from randomized trials using external data: a review of methods. Med Decis Mak. 2017;37(4):377-90.
11. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. The Extrapolation Performance of Survival Models for Data With a Cure Fraction: A Simulation Study. Value in Health. 2021;24(11):9.
12. Rutherford MJ, Lambert PC, Sweeting MJ, Pennington B, Crowther MJ, Abrams KR, et al. NICE DSU Technical Support Document 21: Flexible Methods for Survival Analysis. http://nicedsu.org.uk/wp-content/uploads/2020/11/NICE-DSU-Flex-Surv-TSD-21_Final_alt_text.pdf; 2020.
13. van Oostrum I, Ouwens M, Remiro-Azócar A, Baio G, Postma MJ, Buskens E, et al. Comparison of Parametric Survival Extrapolation Approaches Incorporating General Population Mortality for Adequate Health Technology Assessment of New Oncology Drugs. Value in Health. 2021;24(9):1294-301.
14. Jakobsen LH, Bøgsted M, Clements M. Generalized parametric cure models for relative survival. Biom J. 2020;62(4):989-1011.
15. Jakobsen LH, Andersson TM-L, Biccler JL, El-Galaly TC, Bøgsted M. Estimating the loss of lifetime function using flexible parametric relative survival models. BMC medical research methodology. 2019;19(1):23.
16. Andersson TML, Dickman PW, Eloranta S, Lambe M, Lambert PC. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. Statistics in medicine. 2013;32(30):5286-300.

17. Jackson CH. flexsurv: A Platform for Parametric Survival Modeling in R. Journal of Statistical Software. 2016;70(8):1-33.
18. Bristol Myers Squibb Pharmaceuticals Ltd. Single technology appraisal: Nivolumab for previously treated locally advanced or metastatic squamous non small cell lung cancer [ID811]: Company evidence submission. https://www.nice.org.uk/guidance/TA483; 2015.
19. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WE, Poddubskaya E, et al. Nivolumab versus docetaxel in advanced squamous-cell non–small-cell lung cancer. New England Journal of Medicine. 2015;373(2):123-35.
20. Latimer NR. NICE Decision Support Unit Technical Support Document 14. Survival Analysis For Economic Evaluations Alongside Clinical Trials - Extrapolation with Patient-Level Data. London: National Institute for Health and Care Excellence (NICE) unless otherwise stated. All rights reserved.; 2013.
21. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in medicine. 2002;21(15):2175-97.
22. Liverpool Reviews and Implementation Group. Single technology appraisal: Nivolumab for previously treated locally advanced or metastatic squamous non small cell lung cancer [ID811]: Evidence Review Group report. https://www.nice.org.uk/guidance/TA483; 2015.
23. Bristol Myers Squibb Pharmaceuticals Ltd. BMS response to the Appraisal Consultation Document (ACD) for nivolumab for previously treated locally advanced or metastatic squamous non-small-cell lung cancer (NSCLC). https://www.nice.org.uk/guidance/TA483; 2016.
24. Liverpool Reviews and Implementation Group. Bristol-Myers Squibb (BMS) response to the Appraisal Consultation Document (ACD) for nivolumab for previously treated locally advanced or metastatic squamous non-small cell lung cancer (NSCLC). Evidence Review Group (ERG) commentary on issues raised. https://www.nice.org.uk/guidance/TA483; 2016.
25. Mitchell M. Engauge digitizer 2019 [Available from: https://markummitchell.github.io/engauge-digitizer/.
26. Baio G. survHE: Survival analysis for health economic evaluation and cost-effectiveness modelling. Journal of Statistical Software. 2020;Accepted for publication.
27. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC medical research methodology. 2012;12(1):9.
28. Burnham KP, Anderson D. Model selection and multi-model inference. A Pratical informatio-theoric approch Sringer. 2003;1229.
29. University of California BU, Max Planck Institute for Demographic Research G. Human Mortality Database. 2019 [Available from: www.mortality.org.
30. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. Comparing current and emerging practice models for the extrapolation of survival data: a simulation study and case-study. BMC medical research methodology. 2021;21(1):1-11.
31. Alarid-Escudero F, Kuntz KM. Potential Bias Associated with Modeling the Effectiveness of Healthcare Interventions in Reducing Mortality Using an Overall Hazard Ratio. PharmacoEconomics. 2020;38(3):285-96.
32. Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies. Cancer Epidemiology. 2012;36(1):16-21.
33. Bullement A, Kearns B. Incorporating external trial data to improve survival extrapolations: a pilot study of the COU-AA-301 trial. Health Services and Outcomes Research Methodology. 2022:1-15.

Tables

Table 1: Cost-effectiveness estimates from different extrapolation approaches

	Absolute Value		Incremental values		ICER
	QALYs	**Cost**	**QALYs**	**Cost**	**(per QALY)**
Replicated submission (no cap)					
Nivolumab	1.29	£85,882	0.74	£65,470	£87,926
Docetaxel	0.55	£20,413			
Replicated submission (with cap)					
Nivolumab	0.95	£72,943	0.39	£54,412	£139,958
Docetaxel	0.56	£18,530			
Replicated ERG approach					
Nivolumab	0.66	£56,985	0.33	£40,799	£124,807
Docetaxel	0.34	£16,186			
Dynamic survival models					
Local trend					
Nivolumab	1.06	£75,060	0.50	£56,699	£113,170
Docetaxel	0.56	£18,361			
Damped trend					
Nivolumab	0.87	£67,328	0.35	£49,600	£141,236
Docetaxel	0.52	£17,728			
Dynamic relative survival models					
Local level					
Nivolumab	0.88	£67,880	0.36	£50,229	£139,657

Docetaxel	0.52	£17,651			
Local trend					
Nivolumab	0.99	£72,990	0.45	£54,847	£122,328
Docetaxel	0.54	£18,143			
Damped trend					
Nivolumab	0.86	£66,899	0.34	£49,196	£142,825
Docetaxel	0.52	£17,702			

ERG = Evidence review group. ICER = incremental cost-effectiveness ratio = incremental costs / incremental QALYs. QALY = quality-adjusted life-years.

Table 2: Mean squared error and bias values, averaged over time.

Relative survival model	Mean squared error: Mean (95% confidence interval)	Bias: Mean (95% confidence interval)
Log-logistic	0.022 (0.020, 0.023)	0.106 (0.017, 0.195)
Nelson relative survival	0.086 (0.084, 0.089)	0.166 (0.031, 0.300)
Flexible cure model	0.202 (0.194, 0.211)	0.174 (0.039, 0.309)
Trend dynamic survival	0.089 (0.086, 0.092)	0.127 (0.059, 0.194)
Damped dynamic survival	0.122 (0.121, 0.124)	0.176 (0.103, 0.249)

Figures

Figure 1: Estimates of the trend in the hazard function from two dynamic survival models

[Footnote: Solid-blue line: point-estimates, with 95% confidence intervals in pale blue. Black line = no trend.]

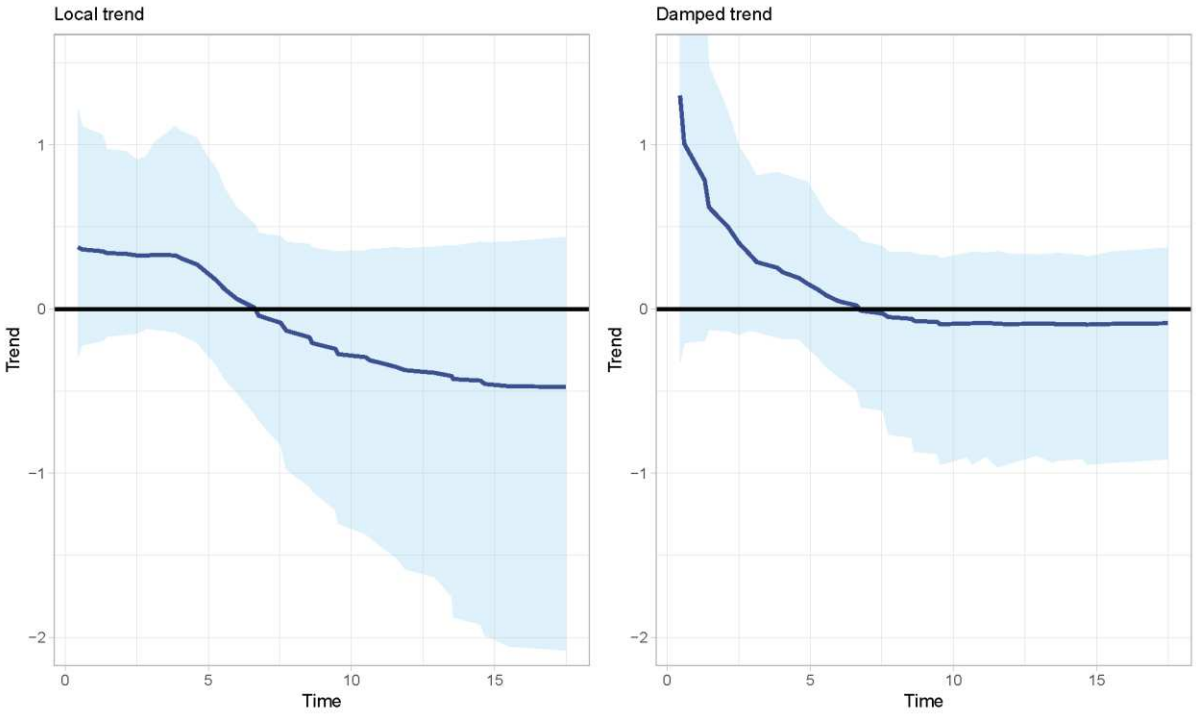


Figure 2: Hazard estimates without external data. Left: within-sample, right: extrapolations

[Footnote: Black line: observed hazard. Red line: general population hazard.]

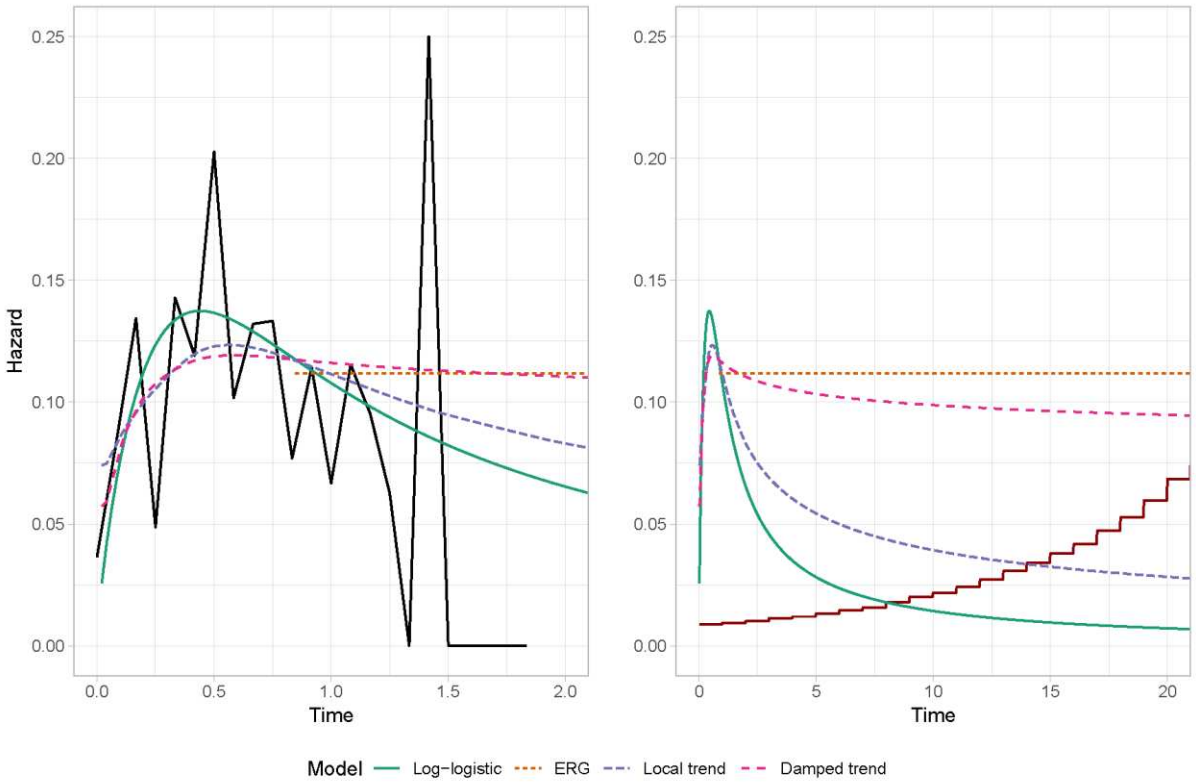


Figure 3: Hazard estimates with external data. Left: within-sample, right: extrapolations

[Footnote: Black line: observed hazard. Red line: general population hazard.]

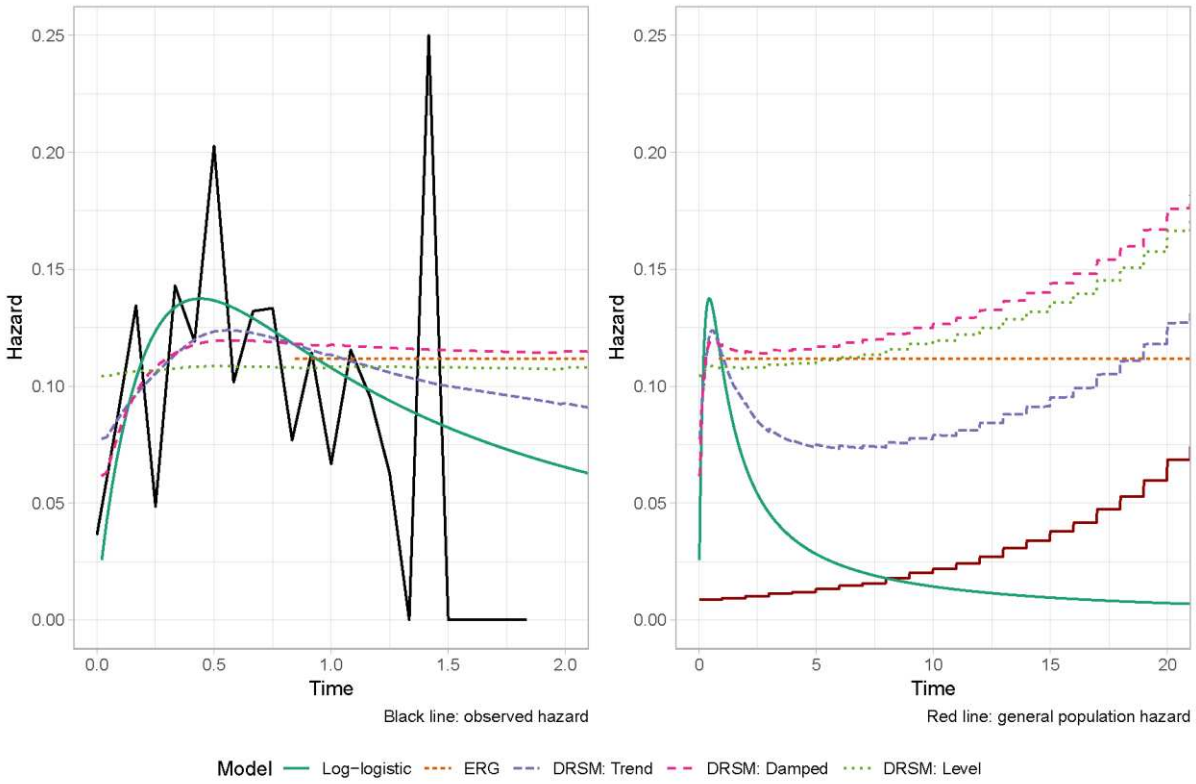
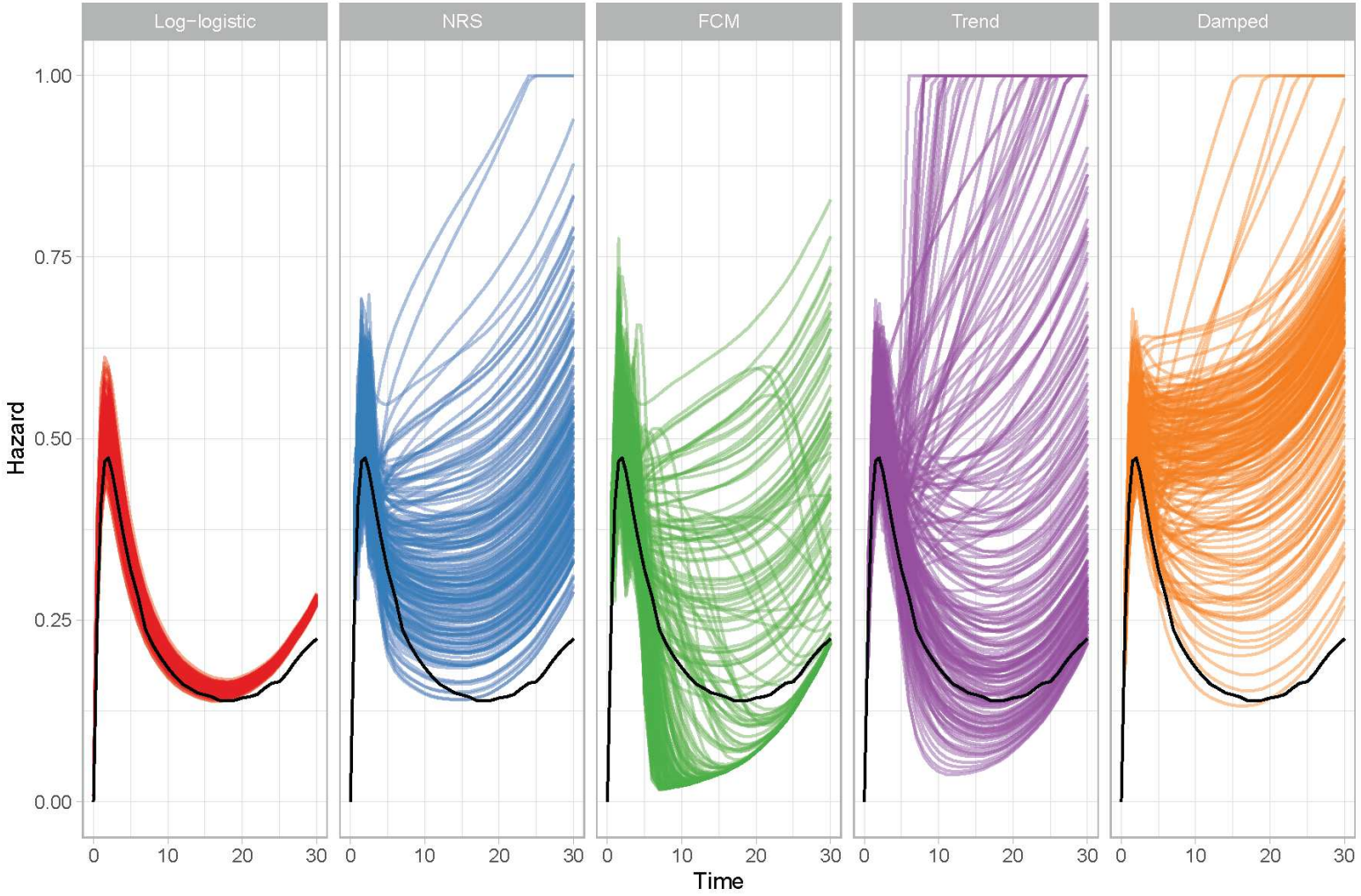


Figure 4: Relative survival model estimates of the hazard function and true values (black lines)



FCM: Flexible cure model. NRS: Nelson relative survival.