Article:
Bullement, A. and Kearns, B. (2022) Incorporating external trial data to improve survival extrapolations: a pilot study of the COU-AA-301 trial. Health Services and Outcomes Research Methodology, 22. pp. 317-331. ISSN: 1387-3741

https://doi.org/10.1007/s10742-021-00264-6

Health Services and Outcomes Research Methodology
https://doi.org/10.1007/s10742-021-00264-6

Incorporating external trial data to improve survival extrapolations: a pilot study of the COU-AA-301 trial

Ash Bullement[1,2] · **Benjamin Kearns**[1]

Received: 3 May 2021 / Revised: 6 November 2021 / Accepted: 10 November 2021

Abstract
Survival extrapolation plays a key role within cost effectiveness analysis and is often subject to substantial uncertainty. Use of external data to improve extrapolations has been identified as a key research priority. We present findings from a pilot study using data from the COU-AA-301 trial of abiraterone acetate for metastatic castration-resistant prostate cancer, to explore how external trial data may be incorporated into survival extrapolations. External trial data were identified via a targeted search of technology assessment reports. Four methods using external data were compared to simple parametric models (SPMs): informal reference to external data to select appropriate SPMs, piecewise models with, and without, hazard ratio adjustment, and Bayesian models fitted with a prior on the shape parameter(s). Survival and hazard plots were compared, and summary metrics (point estimate accuracy and restricted mean survival time) were calculated. Without consideration of external data, several SPMs may have been selected as the 'best-fitting' model. The range of survival probability estimates was generally reduced when external data were included in model estimation, and external hazard plots aided model selection. Different methods yielded varied results, even with the same data source, highlighting potential issues when integrating external trial data within model estimation. By using external trial data, the most (in) appropriate models may be more easily identified. However, benefits of using external data are contingent upon their applicability to the research question, and the choice of method can have a large impact on extrapolations.

Keywords Survival analysis · External data · Extrapolation · Cost-effectiveness analysis

✉ Ash Bullement
abullement1@sheffield.ac.uk

Benjamin Kearns
b.kearns@sheffield.ac.uk

[1] School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

[2] Delta Hat, Nottingham, UK

Published online: 11 January 2022

1 Introduction

Extrapolations of future overall survival (OS) are a key component of many economic evaluations in health technology assessment (HTA). Due to an absence of data (especially in the longer term), extrapolations are inherently uncertain. Depending on the choice of OS extrapolation, a given intervention may be considered cost effective when evaluated as part of an economic model (and therefore be reimbursed), or potentially not be reimbursed if not deemed cost effective.

Often when undertaking survival extrapolation of treatment effectiveness for HTA, the main source of evidence comes from its pivotal clinical trial(s); however, other sources of evidence may help with extrapolations. A better understanding of how these *external* data may be used to improve extrapolations has been identified as a key research priority, and is of particular interest where standard approaches that do not make use of external data yield unrealistic OS extrapolations (Jackson et al. 2017).

In this article, we consider a case study of abiraterone acetate (AA, Zytiga®, Janssen-Cilag Ltd) for castration-resistant metastatic prostate cancer previously treated with a docetaxel-containing regimen. AA was the subject of a National Institute for Health and Care Excellence (NICE) technology appraisal, for which the pivotal COU-AA-301 trial provided the key clinical evidence to support decision making related to both the clinical and cost effectiveness of AA (National Institute for Health and Care Excellence (NICE) 2012).

Two published COU-AA-301 analyses are available: an interim analysis by de Bono et al. (median follow-up of 12.8 months) and a final analysis by Fizazi et al. (median follow-up of 20.2 months) (de Bono et al. 2011; Fizazi et al. 2012). At the time of submission to NICE, extrapolations of lifetime OS based on COU-AA-301 were uncertain, and so any extrapolations from these data were subject to several limitations.

While the COU-AA-301 data are limited, a number of other trials have been conducted in a metastatic prostate cancer population, which may provide useful information when determining appropriate OS extrapolations (Dellis et al. 2019). While absolute values of OS may vary across trials (due to differences in patient characteristics and treatments received), trajectories (shapes) of survival curves may be similar (an assumption commonly made within the context of conducting a network meta-analysis). For example, trial inclusion criteria can lead to an artificial suppression of mortality, followed by increasing mortality due to death amongst more frail patients and then a potential decrease to the lower mortality rates of less frail patients. Other trial data in the same disease area can be obtained from the published literature, with patient-level data recreated using existing algorithms (Guyot et al. 2012). Though some differences in patient populations and intervention effects are expected, it is anticipated that there are likely similarities across trials in patterns of survival over time.

Using the COU-AA-301 trial as a motivating example, this research aims to explore if extrapolations could be improved by combining COU-AA-301 and external trial data. More specifically, the research was focused on the feasibility of using external data within the context of HTA decision making.

2 Methods

In this section, the approach taken to identify suitable external data is described, along with how these data were used to inform survival projections. The section concludes with a description of how survival projections compare when external data are included using different methods, using a range of quantitative and qualitative approaches.

2.1 Identification of suitable external sources

The NICE website was searched for published technology appraisal guidance for prostate cancer interventions using the search term "prostate". For each appraisal, the pivotal trial used to inform the manufacturer's economic analysis was identified via final NICE guidance. Using information from NICE guidance, published references providing information on OS from each trial were identified via a targeted search. Data from multiple interim analyses (where reported) were identified.

2.2 Extraction of relevant data

Kaplan–Meier estimates of survival probabilities (KMs) for the outcome of OS were extracted from each source. These curves were digitised using WebPlotDigitizer (v4.3) (Rohatgi 2020). After procurement of the digitised data, pseudo-individual level data were estimated using a published algorithm (Guyot et al. 2012).

The identified pivotal trials were categorised based on cancer stage and line of therapy. In addition, baseline median serum prostate specific antigen and the percentage of patients with bone metastases/disease spread to the bone were extracted—two established prognostic variables in prostate cancer, that may also be used to establish the extent of disease (and thus, the comparability of studies) (Fizazi et al. 2015; Kuriyama et al. 1996). Using these four criteria, suitably similar trials were determined and deemed eligible for consideration in informing the estimation of OS (with explicit rationale for excluding individual studies documented).

2.3 Inspection of survival data

Data from the latest interim analysis from external studies were taken forward. For each trial, the treatment arm was denoted as either "active" or "placebo".[1] An initial comparison of KMs was undertaken to identify any studies that may appear to exhibit different patterns of OS over time. To further understand the pattern of OS over time, hazard plots were also produced to provide an estimate of the instantaneous risk (hazard) of death at each point in time.

[1] Trial reporting was used to determine grouping. One exception was made for mitoxantrone monotherapy, which was included in the "placebo" group. A systematic literature review by Collins et al. found that mitoxantrone + corticosteroids versus corticosteroids was not associated with any statistically significant improvement in OS (Collins et al. 2006).

2.4 Approaches taken to leverage external data within survival estimation

The approaches to fitting survival models with and without external data are described in Fig. 1. The first two approaches have no formal incorporation; with either no consideration of external data at all (1) or informal consideration when determining the most suitable extrapolation (2). The next two approaches adopt a piecewise approach with external data used after a given timepoint, without adjustment (3) or with further adjustment via the specification of a hazard ratio (HR) assuming proportional hazards (4). The final approach (5) considers models fitted within a Bayesian framework, incorporating external evidence via a prior distribution assumed for the shape parameter (method initially proposed by Soikkeli et al.) (Soikkeli et al. 2019). A more detailed explanation of the latter three methods is provided in the supplementary appendix [SA].

All models were fitted using *R* statistical software, via the *flexsurv* and *survHE* packages (Baio 2018; Jackson et al. 2019; R Core Team 2020). The combination of the *flexsurv* and *survHE* packages allows for all models to be fitted within *R*, as well as subsequent analyses to be undertaken in *R* as needed (such as generation of plots and extraction of parameters).

2.5 Comparison of survival projections

To compare the different approaches a range of methods were used, including visual inspection of the fitted survival models (versus the COU-AA-301 final analysis KM) and interpretation of hazard-based plots. Other metrics, such as statistical goodness-of-fit scores (information criteria; Akaike [AIC], Bayesian [BIC] and deviance [DIC]) (Spiegelhalter et al. 2014), point estimate accuracy (i.e., the absolute difference in estimated survival probabilities at a given time point between the KM estimate and the survival model) and restricted mean survival time (RMST) at multiple time-points were also produced for completeness. However, for the purpose of this research, focus was placed on the plausibility of the models based predominantly on the qualitative assessment of the hazard function (and how the external data may be used to guide model fitting; for example, if there is an indication that long-term hazards will be monotonic or have a turning-point).

3 Results

3.1 Exploration of interim analysis data from COU-AA-301

A summary of the interim analysis data from COU-AA-301 is provided in Fig. 2. Two alternative hazard plots were produced based on smoothed and piecewise constant hazard estimates, using the *R* package *muhaz* (Hess and Gentleman 2019). The default settings of this package were used, with the maximum time value aligned with the end of the KM.

Until approximately 1 year of follow-up, there is a clear advantage associated with AA versus placebo. However, after this time, the curves appear to converge, illustrated in the hazard plots by the smoothed curves crossing. Interim analysis data from COU-AA-301 are limited after 1 year (median follow-up approximately 12.8 months). Up to this time-point, the hazard of mortality appears to be increasing for both treatment arms. Owing to their

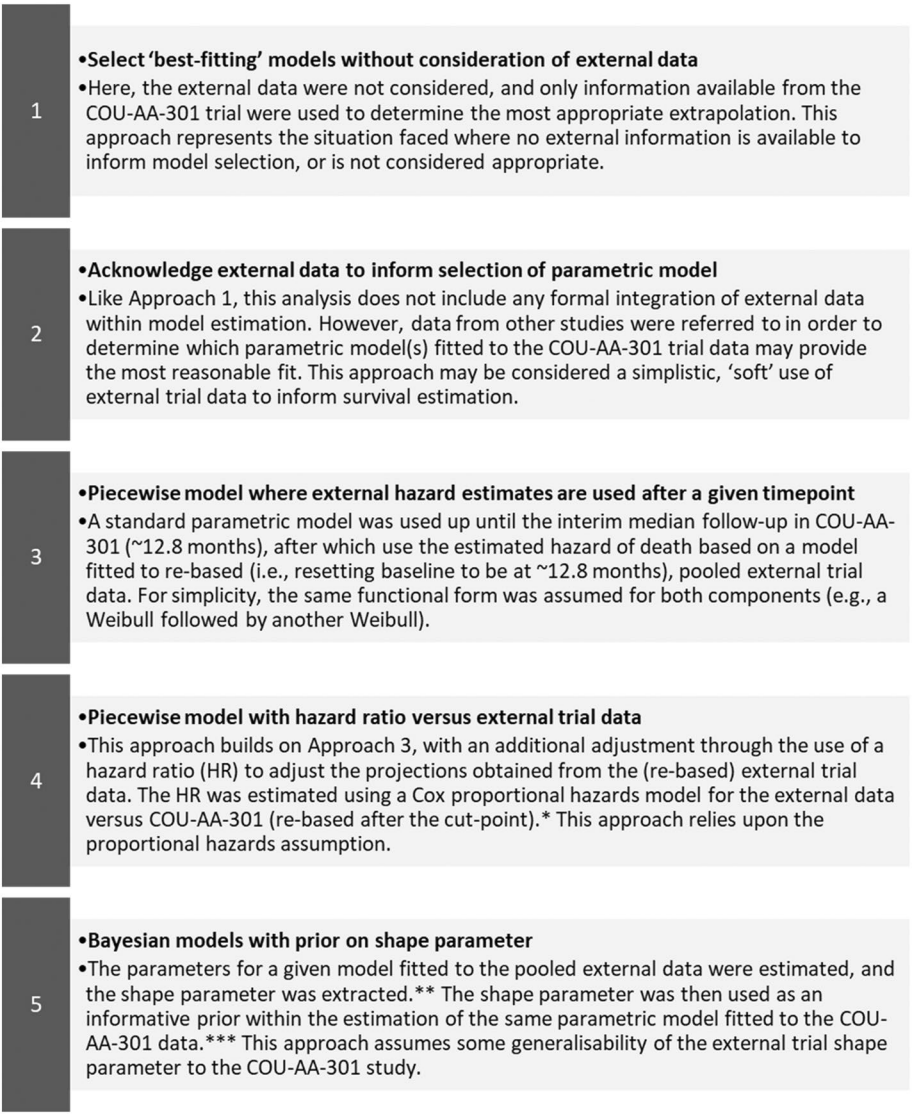

Fig. 1 Summary of modelling fitting approaches. *Depending on the functional form chosen, applying a HR to a base curve may limit the interpretability of treatment effects on specific model covariates. Guyot et al. also highlight that an HR estimated via a Cox PH model will not have the same numerical value as an HR estimated by fitting a parametric model to both arms (Guyot et al. 2011). Despite these limitations, for the purpose of this research, this technical limitation was overlooked but should nevertheless be taken into consideration when interpreting results and considering the use of this method in other contexts. **Note for the generalised gamma model, two shape parameters were extracted. ***For the purpose of this analysis, the shape parameter was extracted and used as the mean of the prior distribution, with a standard deviation (SD) of 5% of the mean value assumed to apply. This imposes a strong prior belief that the 'true' shape parameter of the model fitted to the COU-AA-301 data is similar to that of the external trial data

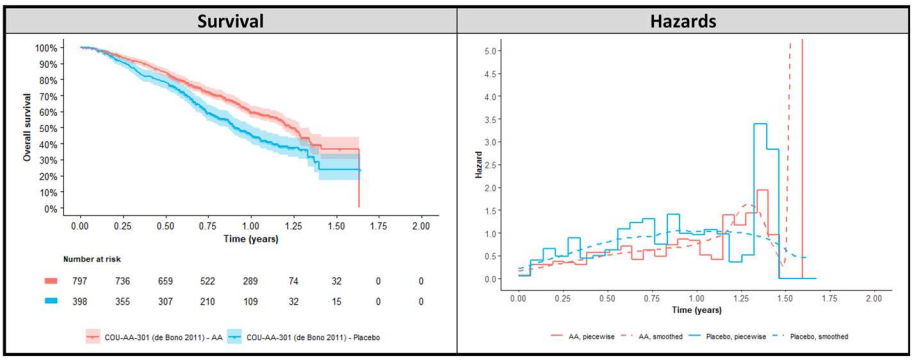

Fig. 2 Survival and hazard plots from interim analysis of COU-AA-301

relatively simpler interpretation, smoothed hazard plots were considered for the remainder of the analysis.

3.2 Identification of suitable external studies

Nine published technology appraisals of prostate cancer treatments were identified (further information provided in SA Table 1). OS data could not be identified for one appraisal which was excluded. Information relating to disease stage, prior use of chemotherapy, serum prostate specific antigen, and presence of bone metastases were collated and compared across trials for the remaining eight appraisals. Subsequently, three studies were not considered sufficiently similar to COU-AA-301 and were excluded (SA Table 2).

A total of five studies were therefore deemed potentially suitable to inform survival modelling. Two studies (PREVAIL and TAX327) were conducted in a chemotherapy-naïve population, whereas the other three studies (AFFIRM, ALSYMPCA, and TROPIC) were conducted in a chemotherapy-experienced/ intolerant population. Importantly, two separate curves are available for enzalutamide, across both a treatment-experienced and a treatment-naïve population (AFFIRM and PREVAIL, respectively).

3.3 Inspection of data from suitable studies

Plots of OS from each of the included studies, including COU-AA-301, are presented in Fig. 3, alongside hazard plots.[2]

Figure 3 shows that outcomes in PREVAIL are noticeably better versus the other studies. This is perhaps because PREVAIL was the only study to have been conducted in a chemotherapy-naïve population, except for TAX327 (the pivotal study of docetaxel, i.e., chemotherapy used prior to the availability of targeted therapies). In addition, the PREVAIL study was published in 2017—the most recent study of those included, whereas the TAX327 study was published in 2008 (i.e., the oldest of the included

[2] Hazard estimates were produced for the maximum follow-up time, and so the ends of the plots should be interpreted with caution. Separate hazard plots were also produced using only OS times after the 12.8 months cut-point (SA Table 1).

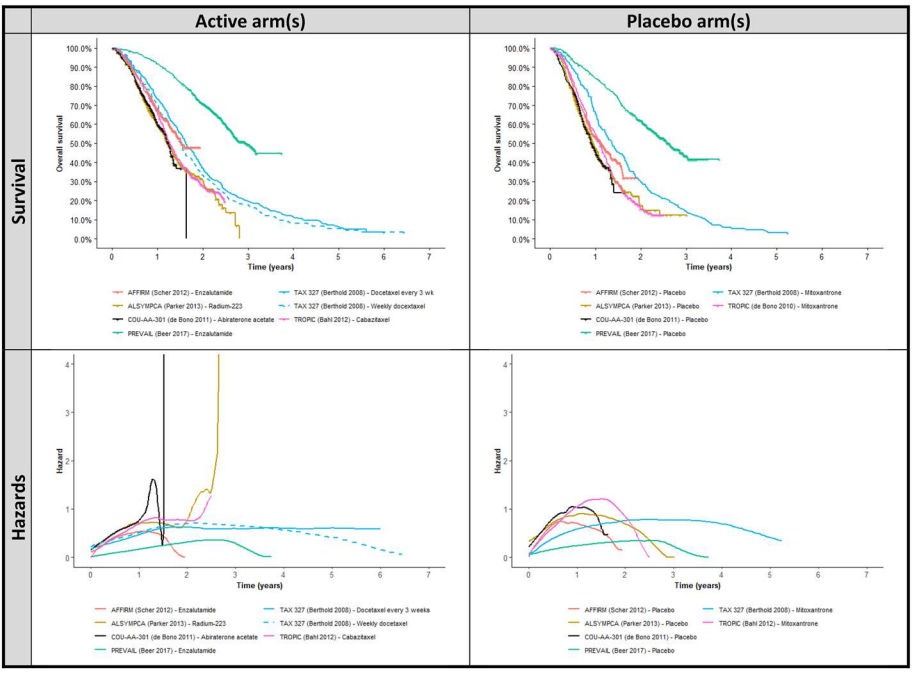

Fig. 3 Survival and hazards from included studies. Note: As noted previously, mitoxantrone monotherapy was designated in the "placebo" group on the basis of a systematic literature review by Collins et al., which found that mitoxantrone + corticosteroids versus corticosteroids was not associated with any statistically significant improvement in OS (Collins et al. 2006)

studies). The longest available follow-up data come from TAX327, wherein the KM estimates extend up to 5.5–6.5 years. All external trials had longer follow-up than the COU-AA-301 interim analysis.

Qualitatively, hazard functions by treatment arm were similar across the trials (as at the start of follow-up, when there is the most data, all functions show an initial increase followed by decreasing hazards), although there is clear variability across the different trials and considerable uncertainty in estimates at the end of follow-up (due to small patient numbers). There is an apparent increase in hazards for at least the first year across the active arms, after which a turning point is seen for some study arms (e.g., enzalutamide in AFFIRM). Conversely, for the placebo arms, hazards appear to decrease by as early as 6 months (e.g., placebo in AFFIRM). All the hazard plots exhibit a non-monotonic pattern, most of which have one turning point. Exceptions to this (e.g., Radium-223 in ALSYMPCA) are likely due to at least one event close to the end of follow-up seen in these two OS curves (see Fig. 3).

The trials with relatively longer follow-up (e.g., TAX327 and PREVAIL) do not exhibit a fall in hazards at the same timepoint as the other studies. This could be due to administrative censoring, small numbers of patients at risk, or differences in study populations across studies.

Owing to the clear differences in PREVAIL versus each of the other studies (including AFFIRM, also a study of enzalutamide), PREVAIL was not considered sufficiently

comparable to inform subsequent modelling. A pooled OS curve (data for the remaining four studies) was considered hereafter.

3.4 Presentation and visual inspection of different approaches

3.4.1 *Simple parametric models (SPMs) with no* use of external data

Simple parametric model (SPM) fits to the COU-AA-301 interim analysis data are provided in Fig. 4. The corresponding hazard plots are provided in SA Fig. 2. In addition, AIC, BIC, and DIC scores for the different models are provided in SA Table 3.

The Weibull SPM may be considered a reasonable fit for either arm based purely on visual fit. Of note, the Weibull provides a very similar fit to the more flexible generalised gamma model. As the Weibull is a special case of the generalised gamma distribution, this suggests that a more complicated model is not required. The Weibull also provided the best fit for the AA arm, versus the log-logistic for the placebo arm. However, several other models provided AIC/BIC scores within 2–3 points of the 'best-fitting' model, and so could also be considered reasonable fits. In addition, the hazard plots (shown in SA Fig. 2) suggest a monotonic model may be inappropriate for longer-term extrapolation.

Other than the exponential, none of the SPMs provided an especially poor fit to the KMs (acknowledging that fit towards the tail end of the KMs may be poor, though this part of the KM should be interpreted with caution). This was confirmed by the similarity of IC values. Therefore, excluding any consideration of external data, several of the SPMs could have been selected as the 'best' fitting.

3.4.2 Simple parametric models (SPMs) with informal use of external data

Figure 4 also compares the SPMs alongside the pooled external data. Comparing the KMs from COU-AA-301 and the external data, the pooled curve follows a similar shape to the COU-AA-301 curve, but outcomes are slightly better for the pooled cohort. Through inspection of Fig. 3 this is unsurprising, given that the curves for the individual studies that inform the pooled curve are associated with either similar or better survival versus COU-AA-301, for both arms.

Based on Fig. 4, the exponential SPM is clearly incapable of fitting to the shape of the KM from either COU-AA-301 or the pooled external data. The Gompertz SPM may be considered to provide a pessimistic extrapolation for both arms, noting that by approximately 3 years nearly all patients are projected to have died (versus > 10% of patients still alive in the external data).

Conversely, the log-logistic and lognormal SPMs provide estimates of OS probabilities for the AA arm that are greater than the external data in the longer term, and are therefore likely optimistic. Corresponding estimates for the placebo arm resemble the external data in the longer term, as the curves fall within the 95% CI limits seen at the tail end of the KM; and therefore, these SPMs may also be optimistic (though more credible than the AA arm fits). Projections for the generalised gamma and Weibull SPMs exhibit a similar shape to the external data.

For the generalised gamma SPM fits to both arms independently, the longer-term hazards (> 2 years) are estimated to be much greater for AA versus placebo (SA Fig. 2). Notably, the generalised gamma SPM estimates consistently increasing hazards in the longer term, even though this model is theoretically capable of reflecting non-monotonic hazards.

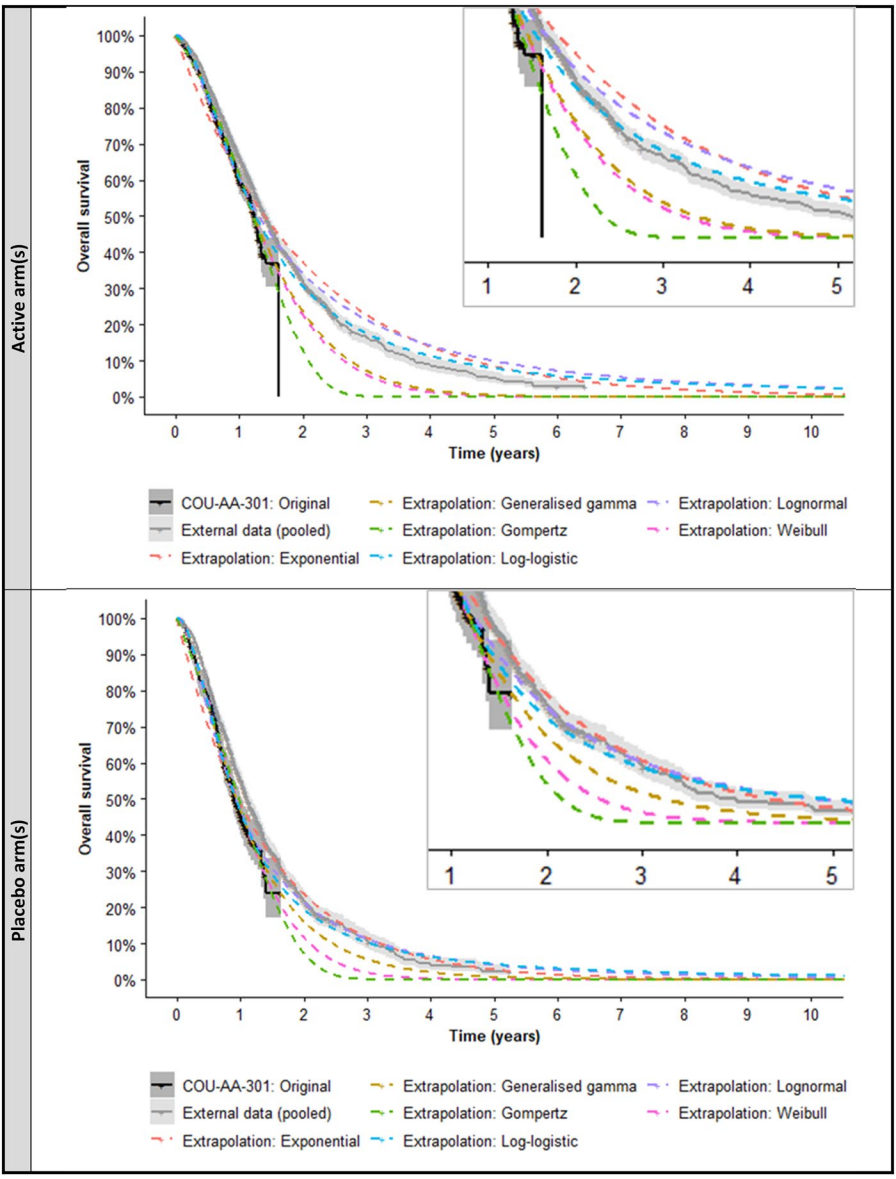

Fig. 4 Simple parametric models fitted to interim analysis compared with external data. Note: Inset plots illustrate difference between fitted models towards the end of the original Kaplan–Meier curve from the original COU-AA-301 data cut

Given that a turning point in the hazard function may be expected, it is likely that the generalised gamma (and other models with monotonic hazards) may under-estimate longer-term OS. Accordingly, the availability of the external data increases our confidence that a model that assumes monotonic hazards (such as the Weibull) is likely inappropriate in this context.

3.4.3 Piecewise models (PWMs)

To produce the piecewise models (PWMs), the same parametric models were fitted to re-based,[3] pooled external data (SA Fig. 3). Plots of the PWMs (alongside corresponding hazard plots) are provided in SA Fig. 4. Sensitivity analyses were also considered with alternative cut points (±3 months either side of the base-case of 12.8 months); however, results were largely unchanged when the cut point was varied, and so are not presented here.

The log-logistic and lognormal PWMs provided notably higher estimates of longer-term OS probabilities, which may be considered unrealistic based on the fit of these models to the rebased external data (SA Fig. 3). Estimates for the remaining four PWMs were largely in agreement. However, due to a combination of there being no adjustments made to the estimated hazards from the external data and the specification of a cut-point where the KMs are close together, the extrapolated tails fit to the external data KM well across both arms. This may be considered unrealistic for the AA arm in particular, as the earlier portion of the KMs suggest improved OS in the external versus COU-AA-301 data, and provides motivation for considering adjustment of the extrapolated portion (discussed further within approach 3).

A further observation is that the PWMs incorporate a turning point in the hazards (SA Fig. 4). This is especially relevant in consideration of the fit of the generalised gamma, Gompertz, and Weibull models; as the SPM fits exhibited a constantly increasing hazard of death, which was misaligned with the estimated external data hazards.

3.4.4 Hazard ratio-based models (HRMs)

Cox PH models were fitted to estimate a HR between the re-based COU-AA-301 data and the pooled external data (separately for the active and placebo arms) (Cox 1972). For the active arms, a HR of 0.675 (95% CI: 0.482, 0.946) was estimated for the external data versus COU-AA-301, indicating a relatively lower hazards for the external data after the cut-point of 12.8 months. For the placebo arms, an equivalent HR of 0.977 (95% CI: 0.586, 1.628) was estimated (external data versus COU-AA-301), indicating little difference in the hazard of death between the sources after the cut-point. Consequently, the HR-based models (HRMs) for the AA arm are expected to produce different estimates compared to the PWMs; however, for the placebo arm, the results are expected to be very similar.

Plots of the HRMs are provided in SA Fig. 5. For the placebo arm, the HRMs yielded very similar results to the PWMs (due to a HR close to 1) and so are not discussed further. For the AA arms, an important difference versus the PWMs is the fact that the extrapolated tails fall close to, or even below, the external data (based on an HR of < 1 for the external versus COU-AA-301 data). The HRMs may therefore be considered to provide a more realistic extrapolation versus the PWMs, notwithstanding the implications of applying a Cox PH-derived HR to a parametric model specified in a non-PH framework.

3.4.5 Bayesian simple parametric models (BSPMs) with prior for shape parameter(s)

Plots of the Bayesian SPMs (BSPMs) are provided in SA Fig. 6. The BPSMs showed generally more consistent extrapolations of OS compared with the SPMs. However, some

[3] See Fig. 1 for explanation of re-based data.

extrapolations still yielded unrealistic longer-term hazards; namely, the Gompertz and Weibull BPSMs which yielded constantly increasing hazards in the longer term for both arms. These models are constrained to have *monotonic* hazards, which the external data suggest is unlikely to be appropriate. Conversely, the lognormal and log-logistic models were shown to produce estimates of longer-term survival probabilities similar to the external data, which (while an improvement on the extent of over-estimation shown for the SPMs) may still be deemed too optimistic. As the exponential model does not have a shape parameter, it could not be considered within this analysis.

3.5 Quantitative comparison of approaches

As described previously, three approaches were considered to compare the different models (statistical goodness-of-fit scores, point-estimate accuracy, and RMST). For the latter two approaches, the main result was based on the median follow-up in the COU-AA-301 final analysis (20.2 months, or 1.68 years), and so each of the percentages expressed below are indicative of OS evaluated at this time. Results are provided over a range of timepoints in the supplementary appendix, as well as AIC/BIC scores.

In terms of point-estimate accuracy (evaluated at 20.2 months), the most plausible SPMs fitted to the AA arm yielded estimates between 32.2 and 37.4% (point estimate value[4]: 36.4%). The PWM estimates were all > 36.4% (range: 37.7–41.1%), likely due to the lack of adjustment to account for differences in the absolute risk of death in the external trial population. All the HRMs under-estimated (range: 30.7–34.9%), suggesting an over-adjustment. Excluding the Gompertz and Weibull BSPMs (given expectation of non-monotonic hazards), 20.2-month OS was estimated within 1.3% (range: 35.8–37.7%). The most plausible SPMs fitted to the placebo arm under-estimated (range: 25.0–26.9%, point estimate value: 28.5%). The PWM and HRM estimates covered a broader range around the point estimate value (range: 25.9–30.9%); whereas the BSPMs under-estimated (range: 20.8–26.6%). Full results are presented in SA Tables 4 and 5.

Through inspection of RMST, estimates based on each modelling approach generally yielded more consistent estimates when making use of the external data (across both arms), and produced a narrower range of estimates. For example, for the AA arm, the HRMs estimated 10-year RMST to be 1.40–1.79 years across all six models, versus 1.22–2.10 for the SPMs. Despite this narrower range, each of the approaches demonstrates high variation dependent upon the functional form assumed. Full results are presented in SA Table 6 and 7.

4 Discussion

This study demonstrates the feasibility of using external trial data from several different studies to inform extrapolations, based on the COU-AA-301 case study. By incorporating external data either implicitly or explicitly within parametric model fitting, the decision of which model(s) may be the most appropriate for HTA decision making can be aided and

[4] Please note that within the context of this study, the point estimate value refers to the actual trial data which should not be conflated with the true underling survival experience of patients outside the COU-AA-301 trial.

estimates of long-term survival probabilities may be improved. However, this finding is conditional upon the external data source providing 'useful' information concerning the expected pattern of OS for the target study.

After the end of the COU-AA-301 follow-up, estimates were shown to differ greatly between the SPMs versus those that incorporated the external data within their estimation, especially those that initially predicted lower survival probabilities. The fact that the range of estimates in the longer term was narrowed following use of external information is helpful in discounting some models not deemed to provide credible extrapolations. For example, the SPMs estimated OS for the AA arm at 20.2 months between 32.2 and 37.4%, whereas the BSPMs estimated an equivalent range of 35.8–37.7% (point estimate value: 36.4%). External data may also provide insight into the long-term behaviour of the hazard function. For example, a turning point in the hazard was seen for the placebo arm in each of the external trials, and in some of the active arms. This information can help to rule-out models which do not incorporate a turning-point.

A somewhat unexpected finding was the large variation in results based on different approaches to incorporating the external information. Previous research has typically focused on a specific method to incorporate external data, rather than comparing across different methods—for example, Soikkeli et al. presented models with a prior distribution set for the shape parameter(s) (Soikkeli et al. 2019). Our findings highlight a broader issue concerning the use of external data, in that the choice of approach (e.g., parametric function and/or method used to incorporate external data) can have a large effect on results, yet it may be difficult to choose which approach is the most suitable in a given scenario. This problem is exacerbated further by the fact that it is not always possible to rely on standard methods for comparing models when fitted with different data sets such as information criteria, as good within-sample fit is not always predictive of good extrapolations (Kearns et al. 2021). Further research is required into identifying which methods work best and when, and may also consider more flexible survival models (Kearns et al. 2019).

This study shows how hazard plots are helpful in choosing between different models. However, their interpretation for the external data is challenging, given that studies were naïvely pooled which may lead to unusual peaks and troughs related to the period of follow-up for each component study. In addition, the hazard estimates for the external data in the longer-term specifically are informed predominantly by one study (TAK-327), owing to its substantially longer follow-up. Despite these limitations, the hazard plots indicate a similar shape for the first year, yet longer-term hazards appear to plateau before falling towards the end of follow-up. It would be of interest to identify if similar hazard shapes were observed for other cancer trials. Similarity of hazard shapes for other trials would also demonstrate the appropriateness of incorporating external evidence to improve long-term extrapolations in other settings.

It was not possible to easily adjust for differences in patient characteristics within the context of this study, as the analysis makes use of several published studies (thus necessitating the re-creation of pseudo-individual level data). Differences between the external studies and the target study could therefore explain some of the inconsistencies in the results produced. Were individual-level data available for all studies, re-balancing of patient characteristics (e.g., via propensity scoring methods) may increase confidence in extrapolations. However, even without such adjustments being possible, the external data may still reveal helpful information in guiding extrapolations (e.g., identification of expected turning points in the hazard function).

The methods considered within this study are by no means an exhaustive set of approaches that may be taken to incorporate external data within extrapolations and are

limited to those that make specific use of other trial data. An alternative Bayesian analysis (beyond specifying a prior for the shape parameter of the model under consideration) may also address the problem of potential imbalances noted above (to an extent), and while beyond the scope of this pilot study, is an area for future research to expand upon these findings. Such research could also consider if incorporating external evidence helps to appropriately quantify extrapolation uncertainty (Kearns et al. 2020).

In our analysis, we considered external trial data to inform OS estimation. However, other types of external information may be of relevance to inform the most appropriate selection of survival model (or perhaps inform the estimation of OS itself). For example, Cope et al. explored how expert opinion may be incorporated with trial data to extrapolate OS (Cope et al. 2019). Here, expert opinion may be used to validate the assumption of a turning point in the hazard function during the extrapolated phase, as suggested by the external trial data. In another study by Guyot et al., a Bayesian analysis was undertaken to incorporate registry data, general population survival, and clinical opinion within extrapolations based on trial data (Guyot et al. 2017). The best approach(es) to combine external data, clinical opinion, and any other relevant information sources within OS estimation remains an outstanding area for further research; and could be considered within a Bayesian framework.

Another implication of this research is the question of how much data should be integrated within survival estimation versus how much should be reserved for validation. In machine learning, it is relatively commonplace to separate a given data source into training, validation, and/or test sets (Brownlee 2017). However, within the context of estimating OS from trial data, there is no established 'gold standard' approach to selecting other studies for estimation versus validation.

In conclusion, this study provides an illustration of how external data may be incorporated into OS estimation; and that by doing so, the range of extrapolations produced by plausible models was narrowed. Use of external data (either formally or informally) may therefore aid decision makers faced with choosing between a broad range of seemingly-plausible survival models. Further research is required to further investigate how external data (and other information sources) may be robustly incorporated within OS probability estimates.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10742-021-00264-6.

Acknowledgements The authors thank Dr. Nicholas Latimer, Prof. Matthew Stevenson, and Prof. Gianluca Baio for their comments on an early version of the manuscript. The authors also thank Prof. Gianluca Baio for his assistance in using the *survHE* package to fit the Bayesian models described in this study. The HEOM Theme of the National Institute for Health Research CLAHRC Yorkshire and Humber (www.clahrc-yh.nihr.ac.uk) supported Ben Kearns in the preparation of this work. The views expressed are those of the authors, and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health and Social Care.

Author contributions BK developed the study design. AB performed the statistical analysis. BK and AB co-wrote the manuscript.

Funding Financial support for this study was provided in part for author BK by the National Institute for Health Research Doctoral Research Fellowship (DRF-2016-09-119) 'Good practice guidance for the prediction of future outcomes in health technology assessment'. Financial support for this study was provided entirely for author AB by the University of Sheffield ScHARR Research Stimulation funding. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Availability of data and material The data used to inform the analysis presented in this study were obtained from articles already in the public domain.

Code availability The code used to perform the analysis presented in this study may be provided by the corresponding author upon reasonable request.

Declarations

Conflict of interest Author Ash Bullement declares that he has received past consulting fees from some of the manufacturers of therapies discussed within the manuscript for unrelated research. Author Benjamin Kearns declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Baio, G.: survHE: Survival analysis in health economic evaluation. R package version 1.1.2 (2018). https://CRAN.R-project.org/package=survHE

Brownlee, J.: What is the Difference Between Test and Validation Datasets?, https://machinelearningmastery.com/difference-test-validation-datasets/, (2017)

Collins, R., Trowman, R., Norman, G., Light, K., Birtle, A., Fenwick, E., Palmer, S., Riemsma, R.: A systematic review of the effectiveness of docetaxel and mitoxantrone for the treatment of metastatic hormone-refractory prostate cancer. Br. J. Cancer. **95**, 457–462 (2006). https://doi.org/10.1038/sj.bjc.6603287

Cope, S., Ayers, D., Zhang, J., Batt, K., Jansen, J.P.: Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia. BMC Med. Res. Methodol. **19**, 182 (2019). https://doi.org/10.1186/s12874-019-0823-8

Cox, D.R.: Regression models and life-tables. J. r. Stat. Soc. Ser. B Methodol. **34**, 187–220 (1972)

de Bono, J.S., Logothetis, C.J., Molina, A., Fizazi, K., North, S., Chu, L., Chi, K.N., Jones, R.J., Goodman, O.B., Saad, F., Staffurth, J.N., Mainwaring, P., Harland, S., Flaig, T.W., Hutson, T.E., Cheng, T., Patterson, H., Hainsworth, J.D., Ryan, C.J., Sternberg, C.N., Ellard, S.L., Fléchon, A., Saleh, M., Scholz, M., Efstathiou, E., Zivi, A., Bianchini, D., Loriot, Y., Chieffo, N., Kheoh, T., Haqq, C.M., Scher, H.I.: COU-AA-301 investigators: abiraterone and increased survival in metastatic prostate cancer. N. Engl. J. Med. **364**, 1995–2005 (2011). https://doi.org/10.1056/NEJMoa1014618

Dellis, A., Zagouri, F., Liontos, M., Mitropoulos, D., Bamias, A., Papatsoris, A.G.: Management of advanced prostate cancer: a systematic review of existing guidelines and recommendations. Cancer Treat. Rev. **73**, 54–61 (2019). https://doi.org/10.1016/j.ctrv.2018.11.005

Fizazi, K., Scher, H.I., Molina, A., Logothetis, C.J., Chi, K.N., Jones, R.J., Staffurth, J.N., North, S., Vogelzang, N.J., Saad, F., Mainwaring, P., Harland, S., Goodman, O.B., Sternberg, C.N., Li, J.H., Kheoh, T., Haqq, C.M., de Bono, J.S.: COU-AA-301 investigators: abiraterone acetate for treatment of metastatic castration-resistant prostate cancer: final overall survival analysis of the COU-AA-301 randomised, double-blind, placebo-controlled phase 3 study. Lancet Oncol. **13**, 983–992 (2012). https://doi.org/10.1016/S1470-2045(12)70379-0

Fizazi, K., Massard, C., Smith, M., Rader, M., Brown, J., Milecki, P., Shore, N., Oudard, S., Karsh, L., Carducci, M., Damião, R., Wang, H., Ying, W., Goessl, C.: Bone-related parameters are the main prognostic factors for overall survival in men with bone metastases from castration-resistant prostate cancer. Eur. Urol. **68**, 42–50 (2015). https://doi.org/10.1016/j.eururo.2014.10.001

Guyot, P., Welton, N.J., Ouwens, M.J.N.M., Ades, A.E.: Survival time outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res. **14**, 640–646 (2011). https://doi.org/10.1016/j.jval.2011.01.008

Guyot, P., Ades, A., Ouwens, M.J., Welton, N.J.: Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med. Res. Methodol. **12**, 9 (2012). https://doi.org/10.1186/1471-2288-12-9

Guyot, P., Ades, A.E., Beasley, M., Lueza, B., Pignon, J.-P., Welton, N.J.: Extrapolation of survival curves from cancer trials using external information. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. **37**, 353–366 (2017). https://doi.org/10.1177/0272989X16670604

Hess, K., Gentleman, R.: muhaz: Produces a smooth estimate of the hazard function for censored data. R package version 12.6.4 (2019). https://CRAN.R-project.org/package=muhaz

Jackson, C., Stevens, J., Ren, S., Latimer, N., Bojke, L., Manca, A., Sharples, L.: Extrapolating survival from randomized trials using external data: a review of methods. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. **37**, 377–390 (2017). https://doi.org/10.1177/0272989X16639900

Jackson, C., Metcalfe, P., Amdahl, J.: flexsurv: Flexible Parametric Survival and Multi-State Models. R package version 2.1 (2019). https://CRAN.R-project.org/package=flexsurv

Kearns, B., Stevenson, M.D., Triantafyllopoulos, K., Manca, A.: Generalized linear models for flexible parametric modeling of the hazard function. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. **39**, 867–878 (2019). https://doi.org/10.1177/0272989X19873661

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 **38**, 193–204 (2020). https://doi.org/10.1007/s40273-019-00853-x

Kearns, B., Stevenson, M.D., Triantafyllopoulos, K., Manca, A.: The Extrapolation performance of survival models for data with a cure fraction: a simulation study. Value Health. (2021). https://doi.org/10.1016/j.jval.2021.05.009

Kuriyama, M., Obata, K., Miyagawa, Y., Nishikawa, E., Koide, T., Takeda, A., Komeda, Y., Kanbayashi, T., Nakano, M., Miyake, K.: Serum prostate-specific antigen values for the prediction of clinical stage and prognosis in patients with prostate cancer: an analysis of 749 cases. Int. J. Urol. off. J. Jpn. Urol. Assoc. **3**, 462–467 (1996). https://doi.org/10.1111/j.1442-2042.1996.tb00577.x

National Institute for Health and Care Excellence (NICE): TA259: Abiraterone for castration-resistant metastatic prostate cancer previously treated with a docetaxel-containing regimen (2012). https://www.nice.org.uk/guidance/ta259/history

R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020)

Rohatgi, A.: WebPlotDigitizer. Version 4.5. Pacifica, CA (2020). Available at: https://automeris.io/WebPlotDigitizer

Soikkeli, F., Hashim, M., Ouwens, M., Postma, M., Heeg, B.: Extrapolating survival data using historical trial-based a priori distributions. Value Health. **22**, 1012–1017 (2019). https://doi.org/10.1016/j.jval.2019.03.017

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: The deviance information criterion: 12 years on. J. r. Stat. Soc. Ser. B Stat. Methodol. **76**, 485–493 (2014). https://doi.org/10.1111/rssb.12062