PharmacoEconomics (2019) 37:1129–1138
https://doi.org/10.1007/s40273-019-00806-4

ORIGINAL RESEARCH ARTICLE

Estimating Lifetime Benefits Associated with Immuno-Oncology Therapies: Challenges and Approaches for Overall Survival Extrapolations

Mario J. N. M. Ouwens[1,6] · Pralay Mukhopadhyay[2] · Yiduo Zhang[2] · Min Huang[2] · Nicholas Latimer[3] · Andrew Briggs[4,5]

Published online: 18 May 2019

Abstract

Background Standard parametric survival models are commonly used to estimate long-term survival in oncology health technology assessments; however, they can inadequately represent the complex pattern of hazard functions or underlying mechanism of action (MoA) of immuno-oncology (IO) treatments.

Objective The aim of this study was to explore methods for extrapolating overall survival (OS) and provide insights on model selection in the context of the underlying MoA of IO treatments.

Methods Standard parametric, flexible parametric, cure, parametric mixture and landmark models were applied to data from ATLANTIC (NCT02087423; data cut-off [DCO] 3 June 2016). The goodness of fit of each model was compared using the observed survival and hazard functions, together with the plausibility of corresponding model extrapolation beyond the trial period. Extrapolations were compared with updated data from ATLANTIC (DCO 7 November 2017) for validation.

Results A close fit to the observed OS was seen with all models; however, projections beyond the trial period differed. Estimated mean OS differed substantially across models. The cure models provided the best fit for the new DCO.

Conclusions Standard parametric models fitted to the initial ATLANTIC DCO generally underestimated longer-term OS, compared with the later DCO. Cure, parametric mixture and response-based landmark models predicted that larger proportions of patients with metastatic non-small cell lung cancer receiving IO treatments may experience long-term survival, which was more in keeping with the observed data. Further research using more mature OS data for IO treatments is needed.

Key Points for Decision Makers

Despite similar and reasonable fits to the observed Kaplan–Meier curve from the evaluated immuno-oncology (IO) trial, the long-term overall survival extrapolation differed substantially across the various survival models examined.

Cure, parametric mixture and landmark models may better account for the potential mechanism of action of IO treatments, whereby a plateau in long-term survival is observed.

A consistent and scientifically grounded approach to survival extrapolations is required to demonstrate the potential value of IO treatments.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40273-019-00806-4) contains supplementary material, which is available to authorized users.

✉ Mario J. N. M. Ouwens
mario.ouwens@astrazeneca.com

1 AstraZeneca, KC4, Gothenburg, Sweden

2 AstraZeneca, Gaithersburg, MD, USA

3 University of Sheffield, Sheffield, UK

4 Memorial Sloan-Kettering Cancer Center, New York, NY, USA

5 University of Glasgow, Glasgow, UK

6 Pepparedsleden 1, 431 83 Mölndal, Sweden

1 Introduction

In many developed countries, including Australia, Canada, Denmark, The Netherlands, Norway, Sweden and the UK, reimbursement decisions for new interventions rely on cost-effectiveness analyses. In oncology, the analyses involve the estimation of lifetime benefits and costs of different treatment options [1–4]. To estimate lifetime survival, the limited follow-up of clinical trials usually necessitates extrapolation of survival data beyond the trial period. The accuracy of extrapolation depends on selecting the appropriate survival distribution. Guidance exists to help with these modelling choices [5, 6]; however, emerging immuno-oncology (IO) therapies may increase the complexity of underlying hazard functions owing to their unique characteristics, including delayed onset of treatment effects and the potential for long-term survival [7], potentially invalidating the use of simple extrapolation methods.

In metastatic cancer clinical trials, the empirical hazards of death may have a complex shape. For example, the use of inclusion and exclusion criteria may suppress the risk (or hazard) of death in the early part of the trial by selecting patients with a lower risk of death than the general metastatic cancer population [8]; however, given that trial participants have severe disease, the hazard of death is likely to increase after the initial trial period. Delayed treatment effects, particularly those of IO therapies, may cause a decline of the death hazard in the medium term, while in the longer term, competing age-related mortality risk may eventually increase the hazard.

Health technology assessments (HTAs) most commonly use exponential or Weibull models [5, 6]. In exponential models, the hazard remains constant over time, while in Weibull models, hazards either increase or decrease monotonically with the exponential as a special case. Other standard models are similarly restricted by the types of hazard changes that they can capture. For example, Gompertz models can represent hazards that increase or decrease monotonically but assume that the rate of change is exponential [9]. Log-logistic, log-normal and generalised gamma models can represent one change in direction of the hazards but cannot characterise any additional directional changes in hazard [6, 10]. Standard parametric distributions may therefore be insufficient to model long-term survival with novel therapies such as IO agents, which are likely to exhibit more changes in the hazard rate over time if they result in delayed effects and long-term survival. Flexible parametric models [11, 12], cure models (CM), parametric mixture models (PMMs) [13, 14] and response-based landmark models, in which survival is modelled for different groups based on a time point when patients can be categorised by treatment response [15], can characterise more complex hazard functions with turning points and changing slopes. However, the relative strengths and limitations of these approaches have not been thoroughly studied and to date they have not been extensively used in HTAs.

In this paper, we discuss these methods and illustrate their application to the ATLANTIC trial (NCT02087423), a phase II single-arm study of durvalumab for previously treated, advanced non-small cell lung cancer (NSCLC) [16]. Durvalumab, a human immunoglobulin G1 (IgG1) monoclonal antibody (mAb), which blocks programmed death-ligand 1 (PD-L1) binding to PD-1 and CD80, is an approved IO therapy in NSCLC [17].

2 Methods

2.1 The ATLANTIC Study

ATLANTIC (NCT02087423) is a phase II, open-label, single-arm trial of durvalumab in heavily pretreated patients with stage IIIB–IV NSCLC with a World Health Organization performance status of 0 or 1, who had recurrent or progressive disease after at least two prior systemic treatment regimens (including one platinum-based regimen) [16]. The primary endpoint was objective response rate, while secondary endpoints included disease control rate, duration of response, progression-free survival, overall survival (OS) and safety. The primary results were published in 2018 [16].

In this analysis, we applied and evaluated different survival modelling approaches to the OS data in the all-treated population from the ATLANTIC trial. OS was measured from the date of first dose to the date of death from any cause, with patients alive at the time of the analysis censored at the date of last contact. The analyses presented in this manuscript are based on an OS data cut-off (DCO) of 3 June 2016 and the all-treated set ($n = 442$) [16]. At that time, the median duration of follow-up in all-treated patients was 8.0 (interquartile range [IQR] 3.4, 13.9) months. Patients continued to be followed for long-term survival; OS results from a subsequent DCO of 7 November 2017 (median follow-up of all-treated patients was 8.9 [IQR 3.4, 22.5] months) were used as a validation set to illustrate how the extrapolations based on the original OS data performed.

Two different ways to smooth the hazards were used. The first approach, run using the 'muhaz' program in R, uses kernel-based methods to estimate the hazard function from right-censored data. Three types of bandwidth function, three types of boundary correction, and four kernel function shapes can be modelled [18]. The second approach, run using the 'B-spline hazard' program in R, accounts for left truncation, right censoring and possible covariates [19].

B-splines can estimate the hazard shape within a generalised linear mixed-model framework [19, 20]. The approach yields smooth estimates of the hazard/survival functions that have a structure intermediate between strongly parametric and non-parametric models [20].

The two smoothing techniques were run using their default settings; the data were also evaluated using different numbers of knots. In addition, monthly raw hazards were computed. The muhaz method was judged by the authors to display more flexibility than the B spline hazard technique (details in Appendix) and, as such, the decision was taken for the muhaz method (with default settings) to be used, together with the empirical hazards, for the remainder of this paper to provide an illustration of the hazards observed during the trial. Empirical monthly hazards were estimated by interpolating the Kaplan–Meier curve at the end of each month, and were computed using the 'survfit' package in R. Note that at the time of extrapolation, data from the new DCO were not available and thus data on longer-term survival and hazards were not used. In the Results section, we present data from the new DCO for illustrational and model validation purposes.

2.2 Modelling Approaches

2.2.1 Standard Parametric Models

First, a set of standard parametric models [6] were fitted to the OS data in the ATLANTIC trial. Standard models considered were the proportional hazards-based exponential, Weibull and Gompertz, and the accelerated failure time-based log-normal, log-logistic and generalised gamma. The standard models have been routinely used in UK National Institute for Health and Care Excellence appraisals and other HTA submissions [6].

2.2.2 Flexible Parametric Models

Spline-based models [11] are flexible parametric models that are defined in stages by polynomial distributions intersected by 'knots'. At each knot, the modelled hazards are smoothed where the distributions change. In simple cases with zero knots, these models are the same as Weibull, log-logistic or log-normal distributions. The spline approach involves not only choosing the number and positions of knots but also the transformation of the survival percentages to the linear prediction scale. We used the transformation of the survival percentages that related to the log-normal distribution because these produced more favourable Akaike information criterion (AIC) and Bayesian information criterion (BIC) results. Results for other transformations can be obtained from the authors upon request.

The following spline models were used to model the OS data: (1) one-knot spline model with knot at 3 months; (2) one-knot spline model with knot at 1 year; (3) two-knot spline model with knots at 3 months and 1 year; and (4) five-knot spline model with knots at 0.25, 0.5, 0.75, 1.0 and 1.25 years.

2.2.3 Cure Models

CMs were first presented more than 50 years ago [21] and have recently been utilised to model survival of novel cancer therapies, such as IO treatments [14, 22]. IO-based studies across different tumour types, including melanoma, have indicated that survival curves eventually plateau, with a significant proportion of patients experiencing a durable long-term survival benefit [21]. The key feature of a CM is the estimation of the percentage of patients who are deemed 'cured', in addition to the estimation of a parametric survival function for patients who are 'not cured' [14]. The risk of death in the 'cured' population is often assumed to be similar to the background population (depending on how 'cure' is defined), while the risk of death for the non-cured population is a mix of background mortality and excess disease mortality. There are two major types of CMs: 'mixture' (MCM) and 'non-mixture' (nMCM). In an MCM, survival is modelled as a mixture of two groups of patients: those who are cured and those who are not (and who therefore remain at risk for the event). In contrast, in an nMCM, it is assumed that all patients belong to the same group, but that event risk decreases to 0 over time, meaning a non-zero proportion of patients will remain alive/will not experience the event in the long-term, even when followed to infinity.

The survival function of the MCM we assessed can be written as shown in Eq. 1:

$$\begin{aligned}\text{Population survival} &= \text{survival of general population} \\ &\times \left(p_{\text{cured}} + \left(1 - p_{\text{cured}}\right) \times \text{survival}_{\text{uncured}}\right)\end{aligned} \tag{1}$$

In the MCM, we used UK and US age- and sex-adjusted mortality data (2012−2014) as background mortality. In order to capture the potential structural difference between the monotone and more flexible distributions, we chose Weibull distribution and log-normal distribution, respectively, as survival functions for the ATLANTIC OS data. The MCM is fitted by flexsurvcure based on relative survival, i.e. a background mortality rate is taken into account for all-cause mortality. As for the standard extrapolation models, the likelihood is then optimised, resulting in both a cure rate and parameter estimates for shape and scale. Further details about how the MCM was fitted are provided in the electronic supplementary Appendix.

The survival function of the nMCM we assessed can be written as shown in Eq. 2:

$$\text{Population survival} = \text{survival of general population} \times \exp\left[\ln\left(p_{\text{cured}}\right)(1-S)\right] \quad (2)$$

where S is a standard extrapolation distribution that decreases to 0 over time, resulting in a cure percentage of $\exp(\ln(p_{\text{cured}})) = p_{\text{cured}}$. In addition, for nMCM, we modelled relative survival to enable this multiplication with the general population survival, and thus with having a hazard that is at least at the level of the general population mortality rate. Further details about how survival was modelled in the nMCM are provided in the electronic supplementary Appendix.

2.2.4 Parametric Mixture Models

PMMs are also used to capture a heterogeneous population; they are used in the case of two or more distinct groups, where there is no assumption of a 'cure'. The model is used to estimate the survival function for patients in each of the distinct groups. A mixture model with two distinct groups can be presented as shown in Eq. 3:

$$\text{Population survival} = p_1 \times \text{survival}_1 + \left(1-p_1\right) \times \text{survival}_2 \quad (3)$$

where p represents the group with lower mortality (i.e. the first mixture) and $1-p$ represents the group with higher mortality (the second mixture). Two scenarios were evaluated: one in which a mixture of two log-normal distributions were used, and one in which a mixture of two Weibull distributions were used, to fit the ATLANTIC OS data, applying Eq. 3 using Bayesian statistics in RJAGS. We assumed that the second group was largest to let the program converge. In addition, we ran the model once assuming that the second group had the longest survival and once assuming that the second group had the shortest survival. Results showed that the assumption of shortest survival did not make sense and can be obtained upon request. Further methodological details about how the PMM was fitted are provided in the electronic supplementary Appendix.

2.2.5 Response-Based Landmark Models

Another approach investigated was a landmark model. This approach models survival for 'responders' and 'non-responders' separately. The response groups are identified at a predefined response evaluation landmark and according to clinical definitions of response. Subsequent survival is modelled from the landmark point to avoid the bias that responders, by definition, have to survive to the point at which response is assessed.

In the ATLANTIC study, response was first assessed at week 8; therefore, the landmark in the model was also defined to 2 months. Response was categorised as (1) responder (i.e. patients who remained progression-free at 2 months); and (2) non-responder (i.e. patients who progressed).

2.2.6 Mean Overall Survival (OS) Estimates

Mean OS was derived from the area under the curve for each survival model. Restricted means were estimated up to the period of the original trial follow-up, and to the period of the new DCO, such that these could be compared with the area under the observed Kaplan–Meier curves for each DCO. In addition, extended mean OS estimates were calculated for each survival model based on the respective extrapolations beyond the trial follow-up, up to 50 years.

All statistical modelling was implemented using the software package R; analyses for the mixture models were also run in RJAGS.

3 Results

3.1 The ATLANTIC Study

Median OS at the original DCO (3 June 2016) was 9.8 months (95% confidence interval [CI] 8.7–11.3), with a 60% maturity (267 events/442 patients started) (Fig. 1). Median OS at the later DCO of 7 November 2017 remained almost unchanged (9.9 months, 95% CI 8.6–11.9), with 74% maturity (326 events/442 patients started).

The empirical and smoothed annual mortality hazards (muhaz) observed in the data are shown in Fig. 2. The default B spline was analogous for muhaz for the mid-period (electronic supplementary Fig. 1).

3.2 All Models

Figure 3 presents the curve fits against the Kaplan–Meier, with Fig. 3a showing the curves with the best AIC of each group (standard and flexible parametric, CM, PMM and response-based landmark) and Fig. 3b displaying, for each group, the curve for which the survival probability at 3.5 years was closest to the 'unknown' new DCO values. A detailed analysis of each model is presented below.

3.3 Standard Parametric Models

The curve fits for standard parametric models are presented in electronic supplementary Fig. 2a (against Kaplan–Meier)

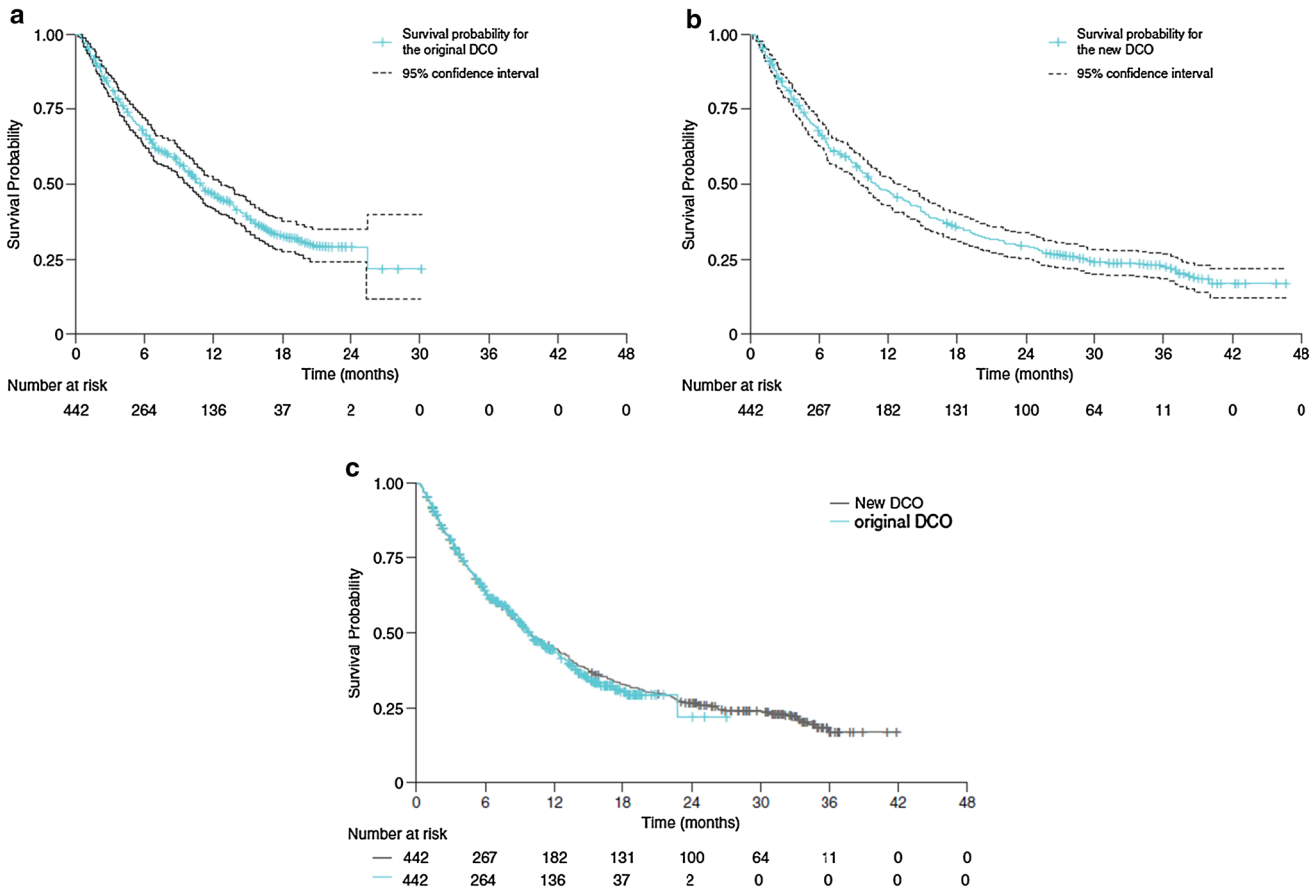

Fig. 1 Kaplan–Meier curve of overall survival in the ATLANTIC study: **a** the DCO used for extrapolations (3 June 2016); **b** the new DCO used for validation (7 November 2017); **c** both DCOs superimposed. *DCO* data cut-off

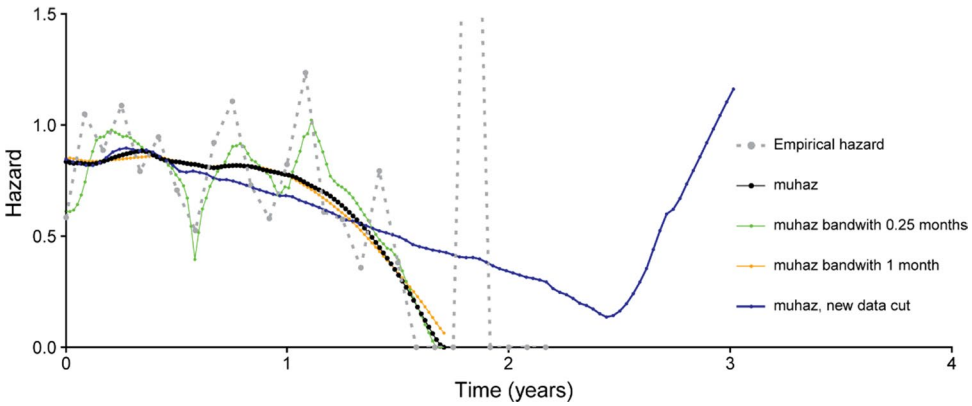

Fig. 2 Hazard plots of the ATLANTIC overall survival data (muhaz)

and Fig. 2b (in terms of hazards). The more complex modelling approaches were compared against the standard log-normal model, which is the best statistical fit among simple parametric distributions according to AIC/BIC. Visual inspection suggested that all standard parametric models provided a reasonable fit to the survival data of the original DCO. However, survival results using the updated DCO showed that all distributions dropped below the observed Kaplan–Meier curve, with only log-normal and generalised gamma curves marginally overlapping the lower bound of the 95% CI.

3.4 Flexible Parametric Models

The fitted curves of the spline-based models shown in electronic supplementary Fig. 2c, d showed an interesting

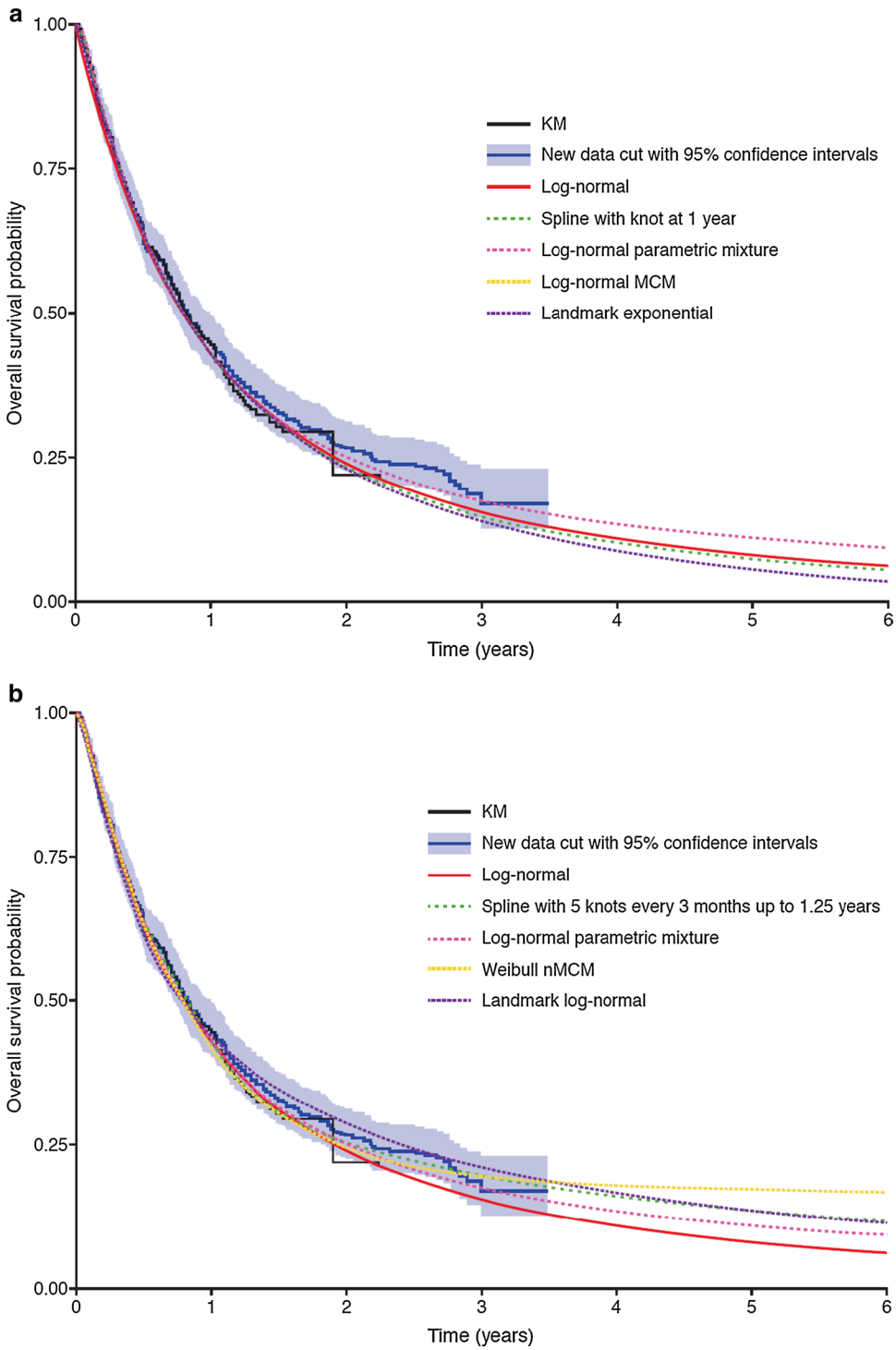

Fig. 3 Curve fits for **a** best Akaike information criterion models against the Kaplan–Meier, and **b** models closest to the new DCO end percentage. *DCO* data cut-off, *KM* Kaplan–Meier, *MCM* mixture cure model, *nMCM* non-mixture cure model

pattern with respect to knot positions. Specifically, models with different numbers of knots predicted similar mortality hazards during the first year. By contrast, the predicted hazards began to differ pronouncedly after 1 year, from which point the Kaplan–Meier became much more uncertain due to low numbers at risk. Different spline models predicted appreciable differences in the survival percentages after the trial period.

3.5 Cure Models

The results for the CM are shown in electronic supplementary Fig. 2e–l. The hazard plots show that CM projected

comparable OS during the trial period compared with standard methods, but they projected improved survival beyond the trial period. Based on background mortality according to the UK life table [23], the log-normal distribution projected statistical cure rates (95% CI) of 0% (0–100) for the MCM and 1% (0–24) for the nMCM. In contrast, the Weibull distribution projected much larger statistical cure rates (95% CI) of 23% (14–33) for the MCM and 19% (8–33) for the nMCM. Use of background mortality based on the US life table [24] rather than the UK life table had minimal impact on the results.

3.6 Parametric Mixture Models

Electronic supplementary Figs. 2m−p provide results from the PMM based on Weibull (electronic supplementary Fig. 2m, n) and log-normal distributions (electronic supplementary Figs. 2o and p). The PMM Weibull distribution produced a much longer tail in the survival curve than the PMM log-normal distribution. For the Weibull distribution, the probability of a patient being in the group with better survival (the first mixture) was 21% (95% CI 0.09–0.29). For the log-normal distribution, the probability was 5% (95% CI 0.00–0.17). Both PMMs suggest the presence of a patient subgroup that achieves a long survival benefit, comparable with the MCM.

3.7 Response-Based Landmark Models

At 2 months, 13% of the ATLANTIC study participants had died and five patients were censored. Of those patients alive at 2 months, 54% were responders (defined as progression-free per Response Evaluation Criteria In Solid Tumors [RECIST] 1.1) and 46% were non-responders. There was a clear differentiation between the two response groups in terms of OS after the landmark point (electronic supplementary Figs. 2q and r).

The extrapolated survival curve for the overall population and the corresponding hazard plot are shown in electronic supplementary Figs. 2s and t, respectively. Inspection of both the survival function and the hazard function suggested that the exponential and Weibull landmark models provided a close fit to the observed data; however, the model with log-normal distribution was more in line with the new DCO.

3.8 Mean OS Estimates for Each Survival Model

Table 1 summarises the mean OS calculated for each survival model. All models provided close in-sample fit to the Kaplan–Meier data, but the out-of-sample fit (Kaplan–Meier+) across the models varied.

4 Discussion

Extrapolation of OS is necessary to estimate lifetime survival based on oncology clinical trials, and is often used to inform cost-effectiveness analyses. Standard parametric survival models are commonly used to estimate long-term survival in oncology HTAs; however, these standard models may not adequately represent the complex pattern of hazard functions, an inadequacy that may further be confounded when evaluating IO treatments owing to their unique characteristics, including delayed treatment effects and potential for long-term survival [7].

We investigated a comprehensive range of models, as alternatives to the standard parametric survival models, to simulate and extrapolate long-term survival outcomes based on the phase II ATLANTIC trial, which evaluated the efficacy of durvalumab in patients with previously treated NSCLC [16]. The availability of updated OS data from this trial presented a unique opportunity to evaluate the performance of extrapolations based on a less mature dataset.

All approaches examined provided a reasonable fit to the observed OS data. However, notable differences were observed when extrapolating beyond the trial period. Therefore, each model led to a different mean OS, which is a key driver in cost-effectiveness models. To identify the most appropriate survival models, the ability to predict long-term survival should be cross-examined with relevant internal and external benchmarks, and be validated against long-term clinical follow-up data (e.g. the later DCO from the ATLANTIC trial) or, if none are available, with real-world data.

We felt that the mixture model (MCM as a special case, if 'statistical cure' can be supported) and the landmark model may better account for the mechanism of action (MoA) of IO therapies compared with standard parametric fitting because they can represent complex hazard functions, which may be observed when treatment effects are delayed and there is long-term survival. These findings are in agreement with those from two recent studies examining the accuracy of standard models for extrapolating long-term OS in IO therapy settings [14, 25]. Both studies found that CMs (or models that included some form of external information, such as registry data) provided much more accurate estimates of long-term OS and associated health economic measures than standard parametric models [14, 25]. In a separate study by Gibson et al., estimates generated by traditional parametric methods were shown to also fit IO progression-free survival K–M curves poorly, whereas those generated from a restricted cubic splines model fitted the curves well [26].

In our analysis, different mixture and CMs projected outcomes differently. For the CM, Weibull and log-normal distributions predicted statistical cure rates of 23% and 0%,

Table 1 Mean OS estimated from each survival model

Model	Section	AIC	BIC	AUC (K-M 2.25; years)	AUC (K-M new 3.5; years)	AUC K-M + (years)	AUC (lifetime; years)
K-M, 2.25 years [95% CI]				1.05 [0.96–1.14]			
K-M new DCO, 3.5 years [95% CI]				1.08 [1.00–1.16]	1.33 [1.21–1.45]		
Minimum across all fitted models				1.03	1.14	0.05	1.19
Maximum across all fitted models				1.11	1.37	9.9	11.22
Weibull	3.3	3793	3801	1.03	1.14	0.05	1.19
Exponential		3791	3795	1.03	1.15	0.07	1.22
Gompertz		3792	3800	1.04	1.20	0.38	1.59
Generalised gamma		3779	3792	1.05	1.25	0.58	1.83
Log-normal		3777	3786	1.05	1.26	0.60	1.86
Log-logistic		3783	3791	1.04	1.24	0.87	2.11
Spline 3 months and 1 year	3.4	3780	3796	1.04	1.21	0.31	1.52
Spline 1 year		3779	3792	1.05	1.25	0.52	1.77
Spline 3 months		3779	3792	1.05	1.26	0.59	1.85
Spline 5 knots 3-month intervals		3783	3812	1.06	1.31	2.00	3.31
MCM log-normal UK	3.5	3772	3785	1.04	1.25	0.51	1.76
MCM log-normal USA		3771	3783	1.04	1.25	0.50	1.75
MCM Weibull UK		3780	3792	1.05	1.34	4.47	5.81
MCM Weibull USA		3779	3791	1.05	1.33	4.28	5.61
nMCM log-normal UK		3773	3786	1.04	1.26	0.84	2.10
nMCM log-normal USA		3772	3784	1.04	1.25	0.84	2.09
nMCM Weibull UK		3779	3791	1.05	1.31	3.71	5.02
nMCM Weibull USA		3777	3790	1.05	1.31	3.60	4.91
PMM Weibull[a]	3.6	3787	NA	1.04	1.32	9.90	11.22
PMM log-normal[a]		3779	NA	1.04	1.28	2.68	3.96
Landmark Gompertz[b]	3.7	2910	2922	1.05	1.23	0.22	1.45
Landmark generalised gamma[b]		2912	2928	1.05	1.23	0.23	1.46
Landmark exponential[b]		2908	2916	1.05	1.23	0.25	1.48
Landmark Weibull[b]		2910	2922	1.05	1.24	0.24	1.48
Landmark log-logistic[b]		2920	2932	1.07	1.32	1.40	2.72
Landmark log-normal[b]		2952	2964	1.09	1.37	1.74	3.11

AIC Akaike information criterion, *AUC* area under the curve, *BIC* Bayesian information criterion, *CI* confidence interval, *DCO* data cut-off, *K–M* Kaplan–Meier, *K-M* + K-M out-of-sample fit, *MCM* mixture cure model, *NA* not available, *nMCM* non-mixture cure model, *PMM* parametric mixture model, *wAIC* Watanabe–Akaike information criterion

[a]wAIC rather than AIC; without background mortality

[b]Landmark AIC/BIC are lower than the other AIC/BIC, because they only assess goodness of fit from a landmark time point onwards

respectively, in the MCM, and 19% and 1%, respectively, in the nMCM. In comparison, for the PMM, the percentages of patients with low mortality were 21% and 5%, respectively. This may suggest some consistency between the CM and PMM approaches but, even with similar fractions in the 'cure' and 'low mortality' groups, these models resulted in substantially different lifetime mean survival estimates (Table 1). We conducted a post hoc nMCM for the second DCO using log-normal and Weibull distributions, and the statistical cure rate was 17% for the Weibull model and 4% for the log-normal model. Similarly, for PMM, the mixture for low mortality was 17% and 6% for Weibull and log-normal, respectively. These results suggest that the cure fractions and mixture fractions for these models were stable over time, but the difference in cure and mixture fractions produced by the different parametric distributions is a cause for concern and led to substantially different long-term mean survival estimates. This is to be expected because, for example, a log-normal model is likely to predict a reducing hazard in the long term, even for the non-cured group, and thus if a cure fraction can be justified, it is likely to be particularly important to consider what the survival distribution in non-cured patients is likely to be.

It appeared that the estimators for the rate of statistical cure and the shape parameter for log-normal distribution were competing for variations in the observed data. As discussed, the standard log-normal distribution may have been able to account for much of the changes seen in the hazard functions; however, the estimator for statistical cure rate may then become redundant. This may also suggest that models that can represent complex hazard functions to some extent, such as the log-normal distribution with its ability to represent a turning point in the hazard function, may make the 'cure rate' difficult to interpret. Therefore, for models with mixture populations, simple distributions, such as Weibull, may be a better choice for the model intuition. At minimum, caution regarding interpretation of the cure rate is needed because it may have different meanings for different survival function distributions. In addition, assigning a distribution that has a long tail to non-cured patients may be unrealistic. Furthermore, it may be reasonable to assume that patients with cancer who are termed 'cured' may still be at higher risk of death compared with the general population. Thus, instead of using a background population in the CM, it may be preferable to use an inflated background population hazard.

The nMCM, PMM and landmark models are designed to address heterogeneous mortality hazards and therefore may be more capable of modelling the potential MoA of IO therapies than spline models or standard parametric approaches, which are more mechanistic. However, to better identify which model provided the best fit and most accurate extrapolation in this setting, more data are required than are currently available in our investigation using ATLANTIC trial data. In addition, it should be noted that even for the models that produced accurate and similar estimates of AUC for the new DCO, their estimates for lifetime mean survival differed greatly.

IO is an emerging therapy area, and the MoA and long-term survival patterns for many agents across different tumour types remain to be fully elucidated, which adds to the challenges of modelling their long-term survival benefits. Improvements in understanding the basic science and MoA of novel IO therapies, coupled with the collection of long-term clinical outcome data, will provide additional information needed to evaluate the appropriate approach for OS extrapolation.

The hazards observed to date in the first DCO from the ATLANTIC trial may not be inconsistent with standard parametric models—they seem to be monotonically decreasing (the increase near the end of the latest data-cut is likely to be highly uncertain). Hence, from the first DCO, we cannot be sure that the investigational drug has resulted in the kind of complex hazard function that we might expect to see with IO treatments.

The second DCO is useful and allows some degree of validation for long-term survival beyond the latest DCO, but caution is required because the data are very limited in order to decide what is reasonable beyond that time point. It seems that several of the models fitted to the original DCO have underestimated survival at the slightly longer time point seen in the new DCO. However, this does not represent conclusive evidence for deciding which is more appropriate beyond that time point out of the flattened survival curves associated with the mixture models, or the more steadily declining survival curves associated with landmark models. There remains a need to revisit this topic with more mature OS data. As a next step, it may also be interesting to perform a simulation study to test which model is the best predictor of long-term OS under a variety of different circumstances.

A limitation of this study is that other biologically plausible combinations of the survival assumptions were not investigated. For example, in the MCM, the cure fraction may be allowed to have a different mortality schedule than the general population. For PMM, background mortality can be included into the estimators, such that the mixture of lower mortality can be more realistic for long-term extrapolations. These other combinations may have had an impact on survival extrapolation outcomes.

5 Conclusions

Despite similar and reasonable fits to the observed Kaplan–Meier curve from the evaluated IO trial, the long-term OS extrapolation from the various survival models differed significantly. This will have a significant impact on cost-effectiveness models and health economic evaluation. The ATLANTIC study showed a slight flattening of the survival curve from the previous to the latest DCO. Standard parametric models fitted to the initial DCO poorly predicted actual survival observed in the later DCO. MCM, PMM and response-based landmark models provided estimates of longer-term survival that were closer to those observed in the later DCO, but themselves resulted in vastly differing estimates of lifetime mean survival.

Even though these models demonstrated theoretical and empirical advantages over standard approaches, it remains a challenge to pinpoint a consistent and scientifically supported approach to extrapolate survival data for IO therapies. Further research using more mature OS data for IO treatments is still needed.

Acknowledgements Medical writing support, which was in accordance with Good Publication Practice (GPP3) guidelines, was provided by

Ewen Buckling, PhD, of Cirrus Communications (Macclesfield, UK), an Ashfield company, and was funded by AstraZeneca.

Author Contributions MJNMO, PM, YZ, MH, NL and AB contributed to the design of the analysis, analysed and interpreted the data, participated in the development of the manuscript, and approved the final draft.

Data Availability Statement Data underlying the findings described in this manuscript may be obtained in accordance with AstraZeneca's data sharing policy described at https://astrazenecagrouptrials.pharmacm.com/ST/Submission/Disclosure.

Compliance with Ethical Standards

Funding The ATLANTIC study (NCT02087423) was funded by AstraZeneca.

Conflict of interest Mario J.N.M. Ouwens and Pralay Mukhopadhyay are employees of AstraZeneca. Yiduo Zhang is an employee of AstraZeneca and owns stock in AstraZeneca. Min Huang is a former employee of AstraZeneca. Nicholas Latimer has received consultancy fees from AstraZeneca in relation to material presented in this manuscript, and has also received consultancy fees from Bristol-Myers Squibb and Pfizer for providing modelling advice. Andrew Briggs has received consultancy fees from AstraZeneca in relation to material presented in this manuscript, and has also received consultancy fees from Bristol-Myers Squibb and Merck (who are manufacturers of immuno-oncologic therapies).

References

1. National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal. London: NICE; 2013.
2. Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. New York: Oxford University Press Inc.; 2006.
3. Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, Krahn M, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. JAMA. 2016;316(10):1093−103.
4. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 4th ed. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2017.
5. Latimer NR. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Making. 2013;33(6):743–54.
6. Latimer N. National Institute for Health and Care Excellence (NICE) Decision Support Unit (DSU) Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data. Report by the NICE DSU, June 2011.
7. Chen TT. Statistical issues and challenges in immune-oncology. J Immunother Cancer. 2013;1:18.
8. Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. Med Decis Making. 2014;34(3):343–51.
9. El-Damcese MA, Mustafa A, El-Desouky B, Mustafa ME. The odd generalized exponential Gompertz distribution. Appl Math. 2015;6:2340–53.
10. Cox C, Chu H, Schneider MF, Munoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Stat Med. 2007;26(23):4352–74.
11. Royston P, Parmar MK. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175–97.
12. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. J Stat Comput Simul. 2015;85(4):777–93.
13. Lambert P. Modeling of the cure fraction in survival studies. Stata J. 2007;7(3):351–75.
14. Othus M, Bansal A, Koepl L, Wagner S, Ramsey S. Accounting for cured patients in cost-effectiveness analysis. Value Health. 2017;20(4):705–9.
15. Latimer N, Ramsey S, Briggs A. Cost-effectiveness models for innovative oncology treatments: how different methodological approaches can be used to estimate the value of novel therapies. International Society for Pharmaceconomics and Outcomes Research 22nd annual international meeting; 20–24 May 2017: Boston, MA.
16. Garassino MC, Cho BC, Kim JH, Mazières J, Vansteenkiste J, Lena H, et al. Durvalumab as third-line or later treatment for advanced non-small-cell lung cancer (ATLANTIC): an open-label, single-arm, phase 2 study. Lancet Oncol. 2018;19(4):521–36.
17. Stewart R, Morrow M, Hammond SA, Mulgrew K, Marcus D, Poon E, et al. Identification and characterization of MEDI4736, an antagonistic anti-PD-L1 monoclonal antibody. Cancer Immunol Res. 2015;3(9):1052–62.
18. CRAN. Hazard function estimation in survival analysis. 2019. https://cran.r-project.org/web/packages/muhaz/muhaz.pdf. Accessed 4 Mar 2019.
19. Rebora P, Salim A, Reilly M. Bshazard: a flexible tool for nonparametric smoothing of the hazard function. The R Journal. 2014;6:114–22.
20. Rosenberg PS. Hazard function estimation using B-splines. Biometrics 199;51:874–87.
21. Berkson J, Gage RP. Survival curve for cancer patients following treatment. J Am Stat Assoc. 1952;47:501–15.
22. Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K, Hamid O, et al. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. J Clin Oncol. 2015;33(17):1889–94.
23. National Life Tables, United Kingdom 2012−2014. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2015-09-23. Accessed 3 Sep 2018.
24. United States Life Tables, 2013. https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_03.pdf. Accessed 3 Sep 2018.
25. Bullement A, Latimer NR, Bell Gorrod H. Survival extrapolation in cancer immunotherapy: a validation-based case study. Value Health. 2019;22(3):276–83.
26. Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, et al. Modelling the survival outcomes of immuno-oncology drugs in economic evaluations: a systematic approach to data analysis and extrapolation. Pharmacoeconomics. 2017;35(12):1257–70.