PharmacoEconomics (2021) 39:345–356
https://doi.org/10.1007/s40273-020-00989-1

ORIGINAL RESEARCH ARTICLE

What Did Time Tell Us? A Comparison and Retrospective Validation of Different Survival Extrapolation Methods for Immuno-Oncologic Therapy in Advanced or Metastatic Renal Cell Carcinoma

Sven L. Klijn[1] · Elisabeth Fenwick[2] · Sonja Kroep[1] · Kasper Johannesen[3] · Bill Malcolm[4] · Murat Kurt[5] · Christopher Kiff[4] · John Borrill[4]

Accepted: 12 December 2020 / Published online: 11 January 2021

Abstract

Background The immuno-oncologic (IO) mechanism of action may lead to an overall survival (OS) hazard that changes over time, producing shapes that standard parametric extrapolation methods may struggle to reflect. Furthermore, selection of the most appropriate extrapolation method for health technology assessment is often based on trial data with limited follow-up.

Objective To examine this problem, we fitted a range of extrapolation methods to patient-level survival data from CheckMate 025 (NCT01668784, CM-025), a phase III trial comparing nivolumab with everolimus for previously treated advanced renal cell carcinoma (aRCC), to assess their predictive accuracy over time.

Methods Six extrapolation methods were examined: standard parametric models, natural cubic splines, piecewise models combining Kaplan–Meier data with an exponential or non-exponential distribution, response-based landmark models, and parametric mixture models. We produced three database locks (DBLs) at minimum follow-ups of 15, 27, and 39 months to align with previously published CM-025 data. A three-step evaluation process was adopted: (1) selection of the distribution family for each method in each of the three DBLs, (2) internal validation comparing extrapolation-based landmark and mean survival with the latest CM-025 dataset (minimum follow-up, 64 months), and (3) external validation of survival projections using clinical expert opinion and long-term follow-up data from other nivolumab studies in aRCC (CheckMate 003 and CheckMate 010).

Results All extrapolation methods, with the exception of mixture models, underestimated landmark and mean OS for nivolumab compared with CM-025 long-term follow-up data. OS estimates for everolimus tended to be more accurate, with four of the six methods providing landmark OS estimates within the 95% confidence interval of observed OS as per the latest dataset. The predictive accuracy of survival extrapolation methods fitted to nivolumab also showed greater variation than for everolimus. The proportional hazards assumption held for all DBLs, and a dependent log-logistic model provided reliable estimates of longer-term survival for both nivolumab and everolimus across the DBLs. Although mixture models and response-based landmark models provided reasonable estimates of OS based on the 39-month DBL, this was not the case for the two earlier DBLs. The piecewise exponential models consistently underestimated OS for both nivolumab and everolimus at clinically meaningful pre-specified landmark time points.

Conclusions This aRCC case study identified marked differences in the predictive accuracy of survival extrapolation methods for nivolumab but less so for everolimus. The dependent log-logistic model did not suffer from overfitting to early DBLs to the same extent as more complex methods. Methods that provide more degrees of freedom may accurately represent survival for IO therapy, particularly if data are more mature or external data are available to inform the long-term extrapolations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40273-020-00989-1.

Extended author information available on the last page of the article

Key Points for Decision Makers

Estimating long-term survival for immuno-oncologic (IO) therapies may require different methods than estimating survival for targeted therapies or chemotherapy.

Piecewise Kaplan–Meier plus exponential models significantly underestimated the survival benefit of IO monotherapy in this case study and should be discouraged for extrapolating IO monotherapy, unless robust contrary evidence for the use of this approach is presented.

In this case study of advanced renal cell carcinoma, a dependent log-logistic model provided accurate estimates of survival across all database locks.

Extrapolation methods with more degrees of freedom may provide accurate estimates of survival if data are more mature or external data can be leveraged to aid model selection.

1 Introduction

Determining the long-term overall survival (OS) benefit offered by novel oncologic therapies is important for health technology assessment (HTA) agencies assessing the cost-effectiveness of these drugs. Given the limited follow-up of many oncology trials at the time of HTA submission, survival extrapolation is usually necessary to assess long-term health benefits and costs. However, the methods used to undertake this extrapolation are a main area of uncertainty for healthcare decision makers and merit further research [1].

This issue is particularly pertinent for immuno-oncologic (IO) therapies. Their mechanisms of action can produce responses that take time to achieve but that can be durable and persist long after treatment has stopped [2, 3]. Hence, the disease-related hazard may show an initial increase, followed by a gradual decline that may tend towards a steady plateau. Given the complex hazard, there is growing evidence that standard parametric extrapolation methods may provide insufficiently accurate estimates of long-term survival for IO therapy [4, 5].

Various alternative survival extrapolation methods are available to capture the IO survival pattern and would be relevant for consideration [6]. However, different methods may yield very different estimates of the long-term survival benefit [1]. Selecting and validating the most appropriate method is thus a key challenge given the importance in determining the cost-effectiveness of these drugs, but current guidelines on this topic are limited [7]. Our study aimed to retrospectively identify the most appropriate extrapolation techniques for an IO monotherapy and a non-IO comparator using long-term follow-up data in advanced or metastatic renal cell carcinoma (aRCC) from CheckMate 025 (CM-025).

2 Methods

We compared and validated six different survival extrapolation methods: standard parametric models (SPM); natural cubic splines (spline); piecewise models combining Kaplan–Meier (KM) data with parametric forms, including the exponential distribution (PieceW-exp and PieceW-other); response-based landmark models (RBLM); and fully parametric mixture models (MM). These methods reflect survival extrapolation recommendations in the National Institute for Health and Care Excellence Decision Support Unit (NICE DSU) Technical Support Document 14 (TSD 14) and those used in recent HTA submissions for IO therapy [8–11]. An overview of the full study design is provided in Fig. 1 and further detailed in the following.

2.1 Data: CheckMate 025

The six extrapolation methods were applied to patient-level survival data from CM-025, a phase III, open-label, randomized controlled trial comparing nivolumab with everolimus for the treatment of aRCC in patients who had received one or two prior lines of antiangiogenic therapy [12]. The primary reason for selecting this dataset was that, at the time of the HTA submissions for nivolumab in pre-treated aRCC, only short-term follow-up data were available (approximately 15 months minimum) [12]. The total survival benefit associated with nivolumab was a main area of uncertainty for HTA agencies, with different agencies using different estimation methods [13–15]. As 64 months' minimum follow-up data from CM-025 have become available (termed herein as '5 year'), this dataset can be used as a case study to compare and validate different extrapolation methods to ascertain those most appropriate for IO therapies [16].

The intention-to-treat population comprised 821 patients, randomized to receive nivolumab 3 mg/kg once every 2 weeks (n = 410) or everolimus 10 mg daily (n = 411). The median OS was 25.0 months (95% confidence interval [CI] 21.8—not estimable) for treatment with nivolumab versus 19.6 months (95% CI 17.6–23.1) for treatment with everolimus. The hazard ratio for death was 0.73 (98.5% CI 0.57–0.93; p = 0.002) for the initial database lock (DBL), which had 15 months' minimum follow-up [12].

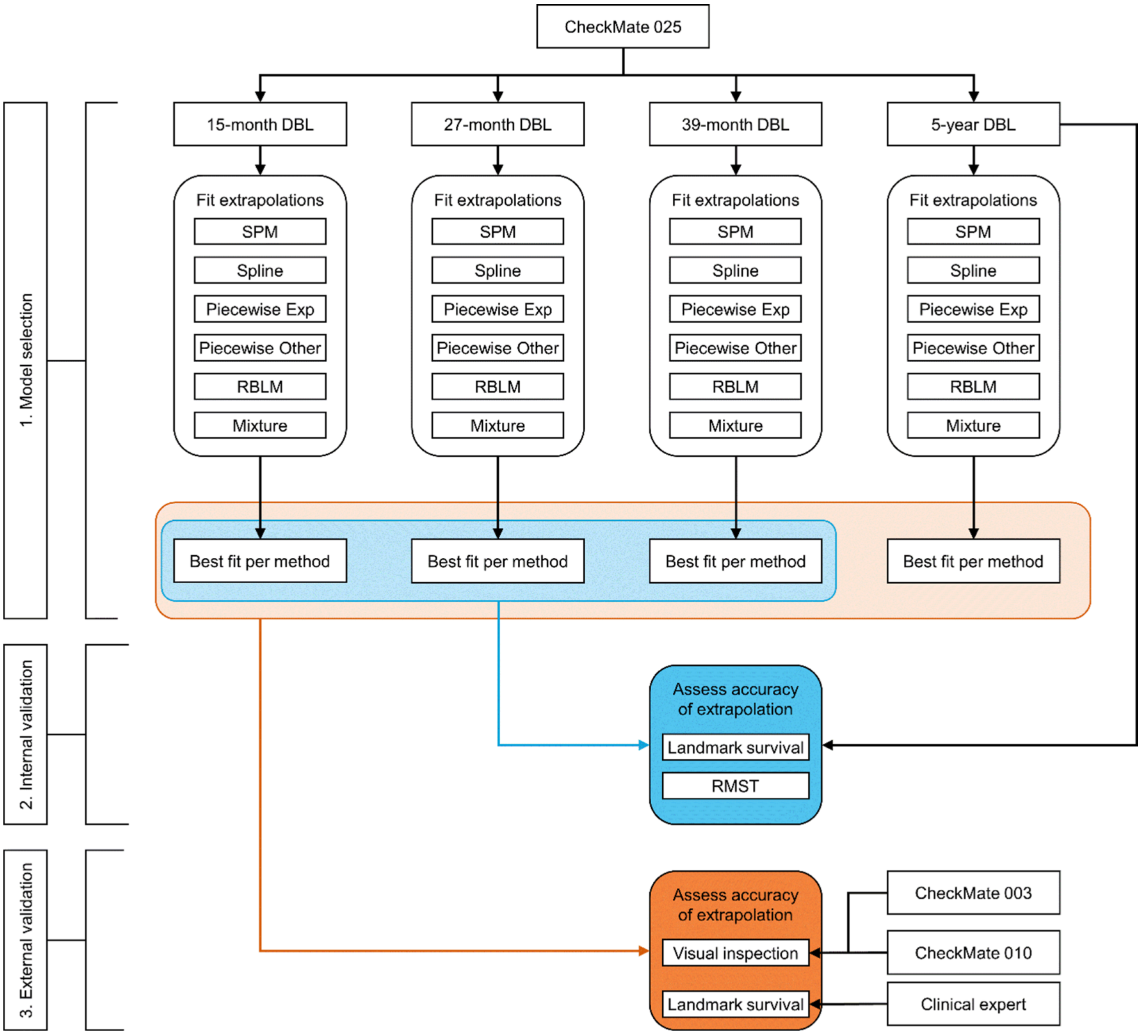

Fig. 1 Study design overview. *DBL* database lock, *Exp* exponential, *RBLM* response-based landmark, *RMST* restricted mean survival time, *SPM* standard parametric model

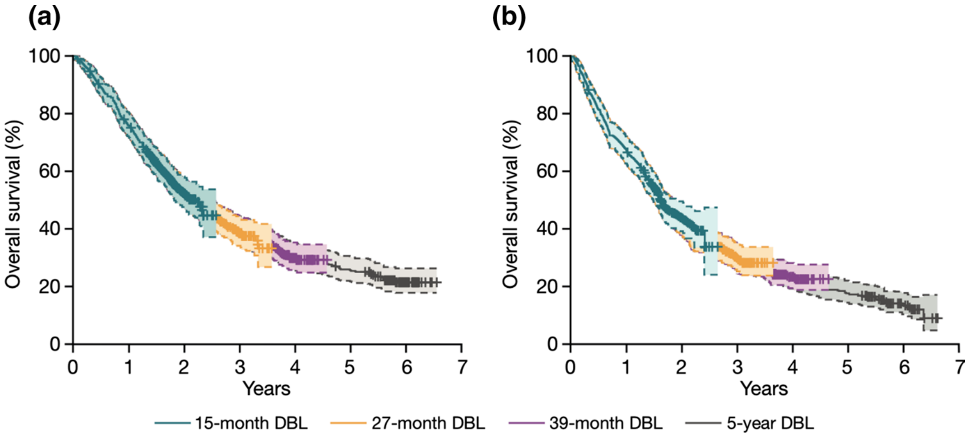

Fig. 2 Overall survival from different database locks (DBLs) in CheckMate 025 for **a** nivolumab and **b** everolimus

Based on the 5-year DBL of CM-025, three DBLs with varying degrees of maturity were created (Fig. 2). Hazard plots can be found in the Electronic Supplementary Material (ESM) (Online Appendix Fig. 1). The first DBL (15 months of follow-up between last subject's randomization and DBL) reflected the initial original trial readout [12]. Two additional DBLs were produced with 1 and 2 years of additional follow-up after the initial DBL (termed '27-month DBL' and '39-month DBL', respectively); these aligned with previously published CM-025 data [17, 18]. These three DBLs were used as a basis for fitting survival extrapolations and are collectively referred to as the 'immature DBLs'.

2.2 Model Selection

Parametric models considered for the SPM were the exponential, gamma, generalized gamma, Gompertz, log-logistic, log-normal, and Weibull. These models were also used within the context of the PieceW, RBLM, and MM methods. For the spline methodology, splines included in the evaluation had either one or two intermediate knots, equally distributed over log uncensored time, using a hazard, odds, or probit model for the transformed survival function, as parametrized by Royston and Parmar [19]. Further details on the methods can be found in the ESM.

For all DBLs evaluated, models were fitted for each of the six different extrapolation methods. Model selection followed NICE DSU recommendations described in TSD 14 [10]. The first step within each DBL involved assessing the proportional hazards (PHs) assumption using visual inspection of the log-cumulative hazard plot, Schoenfeld residuals plot, and a Grambsch–Therneau test [20]. Where the PH assumption could not be rejected, both dependent and independent SPMs and splines were fitted, as independent models could be relevant for scenario analyses. Where the PH assumption was violated, only independent SPMs and splines were fitted. For the PieceW, RBLM, and MM methods, only independent models were fitted.

The second step involved evaluating the fit to the immature DBL to identify the best-fitting model within each method. Models with a poor visual fit to the immature DBL, or that were deemed implausible—for example, those with a hazard below background mortality at 10 years—were not considered for selection. All extrapolations within a method that satisfied visual inspection were ranked according to statistical fit as assessed by Akaike's information criterion (AIC) [21] and Bayesian information criterion (BIC), with BIC taking precedence in case of discordance [22]. The model with the best statistical fit within each method was then used in the internal and external validation steps. In addition, mean lifetime survival was evaluated.

For the PieceW method, the cut-off point for switching from KM data to the parametric extrapolation had to be identified a priori to the evaluation of the fit to each immature DBL. Selection of the cut-off point was empirical and based on visual inspection of the cumulative hazards plot; the inflection point of the cumulative hazard function was used as cut-off point for the KM data.

For the RBLM, a landmark timepoint to evaluate best overall response per *Response Evaluation Criteria in Solid Tumors* v1.1 had to be selected. Selecting a later landmark would capture more responses but yield a less robust model, as OS data after the landmark need to be evaluated separately to avoid immortal time bias. Conversely, selecting an early landmark could under-identify responders. Median time to response in CM-025 was 3.5 and 3.7 months for nivolumab and everolimus, respectively [12]. Therefore, a landmark of 4 months was used in the base-case analyses.

For MMs, several additional criteria were evaluated to select the best-fitting model. First, convergence of the optimization algorithms that maximize the likelihood function to a global optimum instead of a local optimum was required. Second, models estimating the weight of one of the latent subgroups less than 5% were ruled out, as they were considered to be more reflective of the SPM for the predominant group than a true MM.

2.3 Internal Validation

The 5-year minimum follow-up DBL of CM-025 was used for internal validation of the survival extrapolations of each method, using landmark survival and restricted mean survival time (RMST). Landmark survival at 60 and 72 months, as predicted by each method, was compared with OS as observed at these landmark points in the 5-year DBL. RMST, also referred to as area under the curve, was evaluated for the interval between the last events of the 39-month DBL and the 5-year DBL. This interval was the only segment not spanned by any of the three immature DBLs but was observed in the 5-year DBL, albeit including censoring, thus allowing validation of predictions made from the immature DBLs. The limits of the aforementioned time segment for the RMST calculation were 55.0–78.7 months for nivolumab and 55.8–79.4 months for everolimus.

2.4 External Validation

Two external clinical trial datasets including nivolumab in pre-treated aRCC were available: CheckMate 003 (CM-003) and CheckMate 010 (CM-010). Patient baseline characteristics are provided in the ESM. CM-003 was a phase Ib dose-escalation clinical trial in various solid tumours, in which 34 patients with pre-treated aRCC were enrolled [23, 24]. Patients received nivolumab 1.0 or 10.0 mg/kg once every 2 weeks for up to 96 weeks until disease progression, unacceptable toxicity, confirmed complete response, or withdrawal of consent. Median OS was 22.4 months (95% CI 12.5–48.6), and data with a minimum follow-up of 63.9 months were available. CM-010 was a randomized, phase II, blinded, dose-ranging clinical trial in pre-treated aRCC, comparing nivolumab doses of 0.3, 2, and 10 mg/kg every 3 weeks until disease progression or unacceptable toxicity [25]. Pooled 3- and 5-year OS rates across treatment arms were 41% and 34%, respectively [26]. Whilst dosing in CM-003 and CM-010 differed from that in CM-025, datasets from these trials were considered to be useful for external validation given their longer follow-up. Validation of the survival predictions was performed through visual inspection and evaluation of predicted landmark survival relative to the 95% CIs of survival in CM-003 and CM-010.

Further external validation was performed by soliciting input from a clinical expert in the area of aRCC. Based on a published elicitation framework [27], a clinical expert was requested to estimate mean, lower, and upper limits of expected landmark survival for each therapy at 10 and 20 years, as well as expected mean lifetime survival. The expert's answers defined the boundaries of clinical plausibility. Extrapolated landmark and mean survival estimates from each of the methods were compared with these clinically plausible boundaries.

3 Results

3.1 Model Selection

The PH assumption could not be rejected in any of the DBLs. The log-cumulative hazard plots showed curves that initially crossed in the first month but then remained relatively parallel (Fig. 2 in the ESM). This was consistent with the Grambsch–Therneau tests, which showed that the PH assumption could not be rejected in any of the DBLs, with $p > 0.05$ in all cases (Fig. 3 in the ESM). Therefore, both dependent and independent SPMs and splines were fitted.

The best-fitting models per extrapolation method and DBL are displayed in Table 1 and visualized for the 15-month DBL in Fig. 3 (and Fig. 4 of the ESM). Model selection within each method was mostly stable across DBLs. Exceptions included the independent SPMs, splines, and MM, where the best-fitting models deviated between the earlier and later DBLs. Additionally, the best-fitting model for the non-responders group of the RBLM was not stable across DBLs.

3.2 Survival Estimates

Figure 4 presents the mean lifetime survival estimates, with a time horizon of up to 40 years, generated by the best-fitting models for each method for each DBL. Survival estimates were adjusted for background mortality, applying the maximum of estimated mortality rates and general population mortality rates as per UK life tables [28]. Estimates for nivolumab showed an increasing trend with maturity of the DBLs for the independent SPM, independent spline, and PieceW-exp methods. The MM method showed a major decrease in estimated nivolumab lifetime survival from the 39-month DBL onward. The other methods were relatively stable across DBLs for nivolumab. Estimates for everolimus mean lifetime survival were relatively stable across DBLs for all extrapolation methods.

3.3 Internal Validation

All but MM methods underestimated landmark survival for nivolumab at 60 and 72 months as compared with observed survival in the 5-year DBL. The MM methods overestimated survival at those landmark timepoints (Fig. 5a–d). Nivolumab landmark survival estimates with the independent SPM, independent spline, PieceW-exp, and RBLM

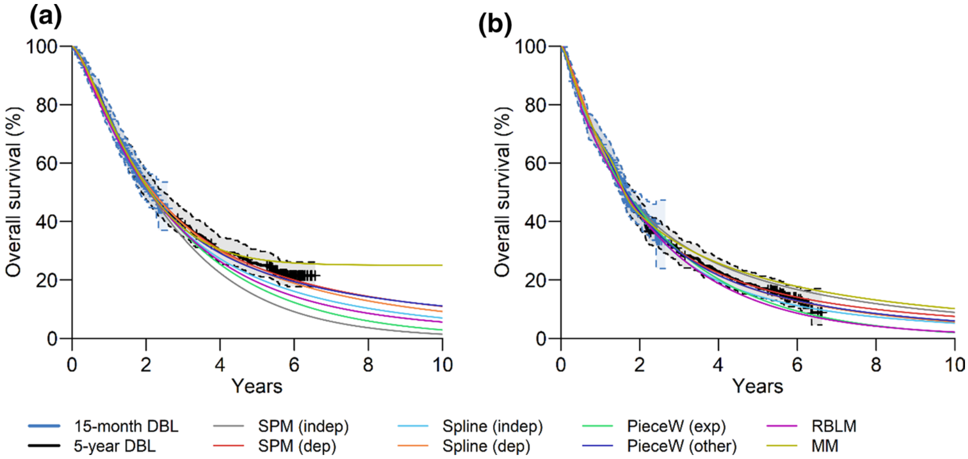

Fig. 3 Best-fitting models for the 15-month DBL for **a** nivolumab and **b** everolimus. *DBL* database lock, *dep* dependent, *exp* exponential, *indep* independent, *MM* mixture model, *PieceW* piecewise, *RBLM* response-based landmark, *SPM* standard parametric model

Table 1 Best-fitting models per method

Method	15-month DBL	27-month DBL	39-month DBL	5-year DBL
Nivolumab				
SPM				
Dependent	Log-logistic	Log-logistic	Log-logistic	Log-logistic
Independent	Gamma	Log-logistic	Log-logistic	Log-logistic
Spline				
Dependent	1-knot normal	1-knot normal	1-knot normal	1-knot normal
Independent	1-knot normal	1-knot odds	1-knot odds	1-knot odds
PieceW				
Exponential	Exponential	Exponential	Exponential	Exponential
PieceW				
Other	Log-logistic	Log-logistic	Log-logistic	Log-logistic
RBLM				
Responders	Log-normal	Log-normal	Log-normal	Log-normal
Non-responders	Exponential	Exponential	Log-logistic	Gamma
MM				
S_1	Exponential	Exponential	Exponential	Exponential
S_2	Weibull	Weibull	Weibull	Gamma
Weight S_1	25.0%	25.7%	42.7%	45.1%
Weight S_2	75.0%	74.3%	57.3%	54.9%
Everolimus				
SPM				
Dependent	Log-logistic	Log-logistic	Log-logistic	Log-logistic
Independent	Log-normal	Log-logistic	Log-logistic	Log-logistic
Spline				
Dependent	1-knot normal	1-knot normal	1-knot normal	1-knot normal
Independent	2-knot odds	1-knot normal	1-knot normal	1-knot normal
PieceW				
Exponential	Exponential	Exponential	Exponential	Exponential
PieceW				
Other	Log-normal	Log-normal	Log-normal	Log-normal
RBLM				
Responders	Exponential	Exponential	Exponential	Exponential
Non-responders	Exponential	Exponential	Log-logistic	Exponential
MM				
S_1	Log-normal	Log-normal	Log-normal	Gamma
S_2	Log-normal	Log-normal	Log-normal	Log-normal
Weight S_1	7.3%	10.1%	9.9%	8.7%
Weight S_2	92.7%	89.9%	90.1%	91.3%

DBL database lock, *MM* mixture model, *PieceW* piecewise, *RBLM* response-based landmark, *SPM* standard parametric model, S_1 latent subgroup 1, S_2 latent subgroup 2

methods, based on the 15-month DBL, were outside the 95% CI of survival as observed in the 5-year DBL. This was also the case with the PieceW-exp and RBLM methods for predictions based on the 27-month DBL. For everolimus, landmark survival estimates were, on average, closer to survival as observed in the 5-year DBL than those for nivolumab. More explicitly, only the PieceW-exp and RBLM estimates based on the 15- and 27-month DBLs fell outside the 95% CI of survival as observed in the 5-year DBL.

RMST for nivolumab was consistently underestimated by all but MM methods, which led to minor overestimations of RMST (Fig. 5e, f). The MM based on the 39-month DBL was the most accurate predictor of RMST for nivolumab. RBLM method estimates fell outside the 95% CI, even if the 27-month DBL was used as a basis for extrapolation. PieceW-exp method estimates fell outside the 95% CI, irrespective of which DBL was used as a basis for extrapolation. RMST estimates for everolimus were, on average, closer to observed RMST than for nivolumab. The SPM, PieceW-other, and MM methods tended to overestimate everolimus RMST, whereas

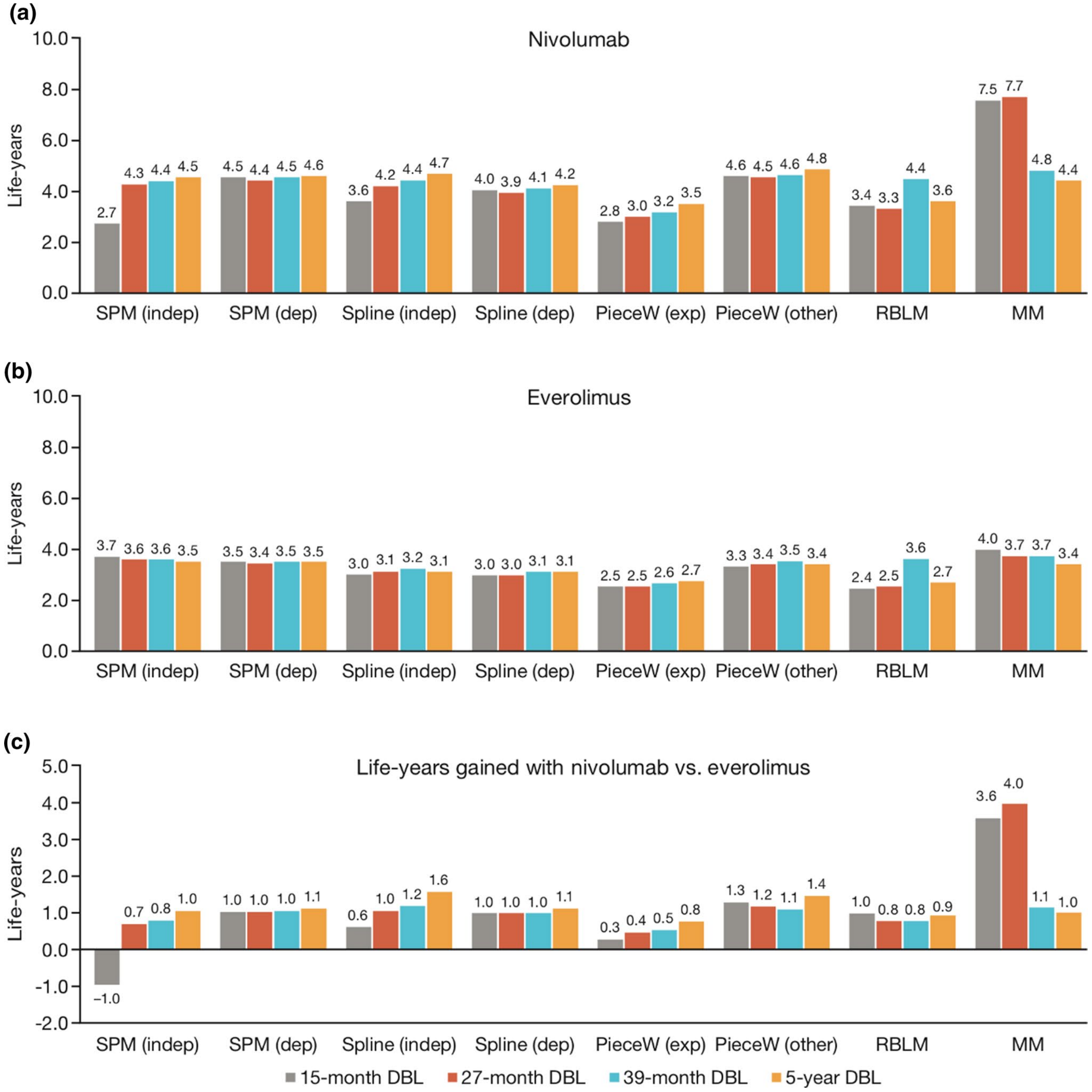

Fig. 4 Mean lifetime survival estimates from the best fits of each extrapolation method for **a** nivolumab, **b** everolimus, and **c** nivolumab relative to everolimus. *DBL* database lock, *dep* dependent, *exp* exponential, *indep* independent, *MM* mixture model, *PieceW* piecewise, *RBLM* response-based landmark, *SPM* standard parametric model

the other methods tended to underestimate RMST. The observed RMST gain of nivolumab relative to everolimus of over 2 months in the 5-year DBL was underestimated by all methods based on the 15-month and 39-month DBLs (Fig. 6).

3.4 External Validation

Comparison with CM-003 demonstrated that most methods, except for dependent SPM and MM, predicted survival for nivolumab to be significantly lower than observed in CM-003, when basing extrapolations on the 15- or 27-month DBLs (see the ESM for a full overview). Comparison with

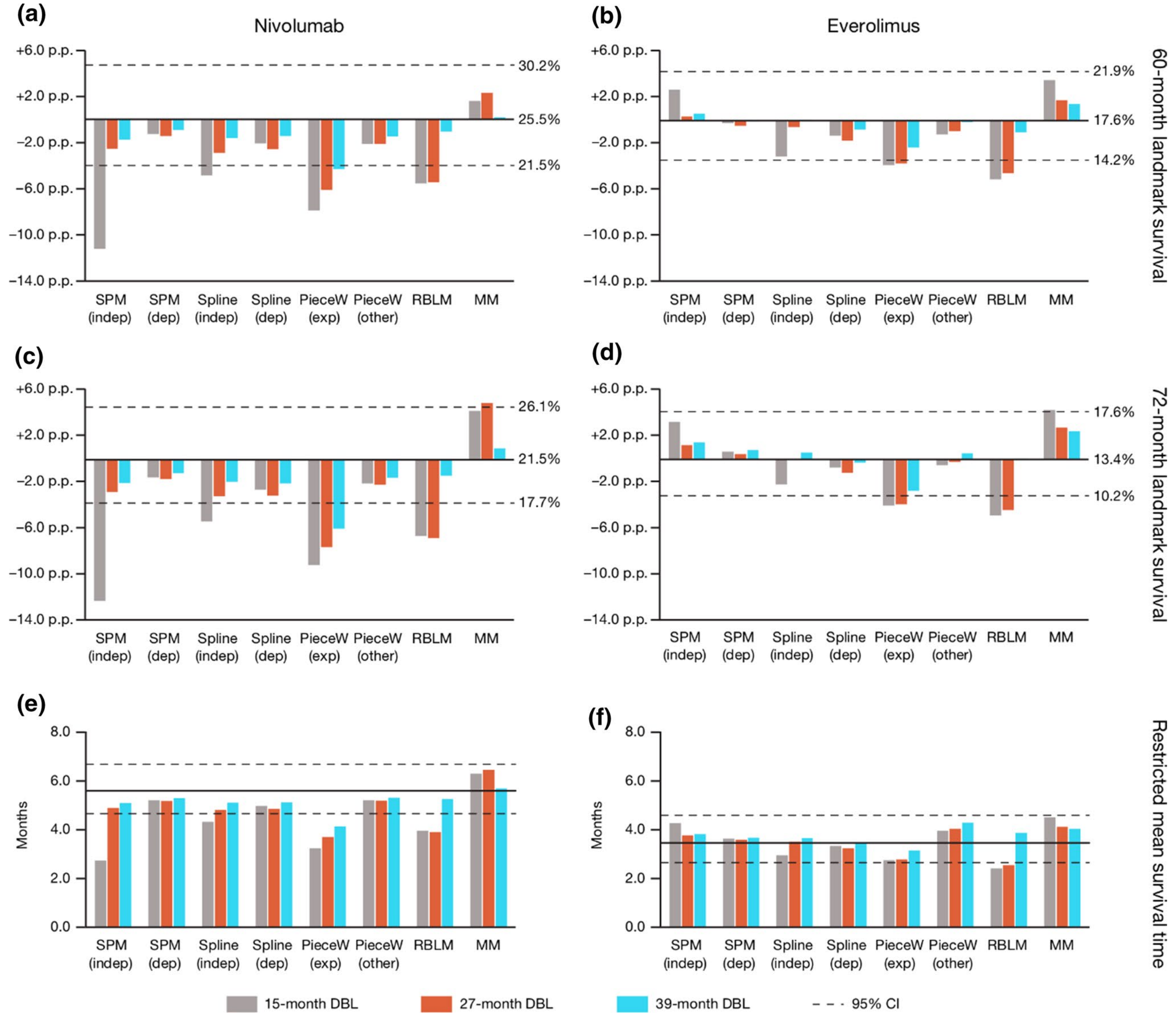

Fig. 5 Comparison of predicted and observed survival (based on 5-year DBL). 60-month landmark survival, deviation from KM data in **a** nivolumab and **b** everolimus, 72-month landmark survival, deviation from KM data, in **c** nivolumab and **d** everolimus, and RMST (solid black line indicates observed RMST, dashed line indicates 95% CI) in **e** nivolumab and **f** everolimus. *CI* confidence interval, *DBL* database lock, *dep* dependent, *exp* exponential, *indep* independent, *KM* Kaplan–Meier, *MM* mixture model, *PieceW* piecewise, *p.p.* percentage point, *RBLM* response-based landmark, *SPM* standard parametric model

CM-010 demonstrated that most extrapolation methods yielded survival estimates within the 95% CI of OS as observed in CM-010, based on the 15- and 27-month DBLs. All estimates from the 39-month DBL fell within the 95% CI of CM-010.

The clinical expert estimated landmark survival for nivolumab to be 15% and 10% at 10 and 20 years, respectively. For the same timepoints, the expert estimated survival for everolimus to be 7% and 5%, respectively. Based on the 15-month DBL, the independent SPM, PieceW-exp, and RBLM all provide survival estimates that fell below the lower limit provided by the clinical expert at 10 and 20 years, respectively. In general, most models underpredicted the estimated nivolumab 20-year survival versus 10-year survival. The MMs on the other hand, overestimated a clinically plausible nivolumab landmark survival based on the 15- and 27-month DBLs. For everolimus, differences were smaller between the methods, though a similar pattern could be detected in which the PieceW and RBLM methods delivered results that could be deemed implausible from the perspective of the clinical expert.

The clinical expert also estimated mean lifetime survival with respective lower and upper limits. For nivolumab, this was estimated as 4 years (lower limit 3 years, upper limit 5

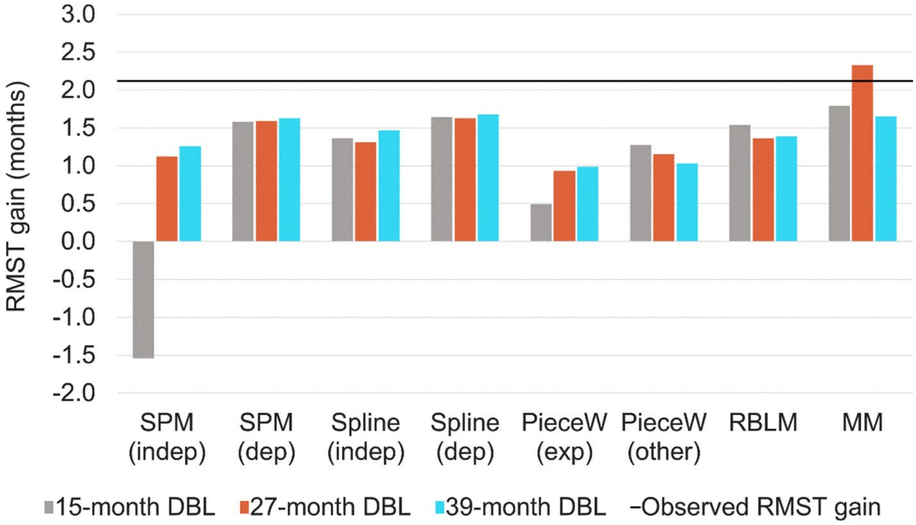

Fig. 6 Restricted mean survival time gain (based on 5-year DBL) and estimations based on the 15-month, 27-month, and 39-month DBLs. *DBL* database lock, *dep* dependent, *indep* independent, *MM* mixture model, *PieceW* piecewise, *RBLM* response-based landmark, *RMST* restricted mean survival time, *SPM* standard parametric model

years) and for everolimus it was 3 years (lower limit 2 years, upper limit 4 years). The independent SPM and PieceW-exp methods provided mean lifetime survival estimates for nivolumab in some cases that were deemed too short to be clinically plausible. The MM method estimates from the 15- or 27-month DBLs could be considered too long to be clinically plausible. For everolimus, it was noticeable that none of the methods provided clinically unreasonable estimates for mean lifetime survival.

4 Discussion

This study revealed that all methods, with the exception of MM, underestimated survival for nivolumab in pre-treated aRCC. The magnitude of underestimation depended on the extrapolation method used and the maturity of the dataset but varied between 2.9 and 0.3 months of underprediction for RMST. Based on the initial 15-month DBL, the independent SPM, independent spline, PieceW-exp, and RBLM methods substantially underestimated nivolumab survival at the 60- and 72-month landmarks, to the point of becoming clinically implausible. Using DBLs with a longer follow-up resulted in lower underestimation.

In contrast, all methods except PieceW-exp and RBLM provided reasonable estimates of everolimus OS. A possible explanation for this observation is that nivolumab and everolimus have notably different mechanisms of action, which may lead to differences in survival patterns over time.

An unexpected finding, given the differences in mechanism of action between nivolumab and everolimus, was that the PH assumption held for all DBLs. A potential explanation may be the use of IO agents as subsequent treatment in patients randomized to everolimus. Because nivolumab was approved by the US FDA and the European Medicines Agency for use in second-line renal cell carcinoma approximately 3 years before the 5-year DBL [29, 30], 26.0% of patients in the everolimus arm received nivolumab as a subsequent therapy or had crossed over to nivolumab in the 5-year DBL. This may be considered reflective of real-world circumstances but may have led to similarities in the tail behaviours of both arms [31].

The dependent log-logistic model performed well across all DBLs. In this case study, it did not suffer from overfitting to short-term trends as much as the more complex methods. The log-logistic distribution can be characterized by an arc-shaped hazard profile, with hazards that increase initially followed by a monotonically decreasing hazard. When implemented in a relative survival framework, accounting for background mortality, the hazard will at some point increase again. Within this case study, the maximum of log-logistic mortality rates and general population mortality rates were applied, leading to a similar hazard profile. Irrespective of the method applied to account for background mortality, the underlying assumption is that patients with aRCC may in the long term have mortality that is similar to that of the general population. Some HTA agencies have previously challenged the use of the log-logistic distribution [14, 32], but our study shows that it seems to be a reasonable candidate for extrapolating long-term survival data for IO monotherapy agents.

The PieceW-exp method performed poorly in predicting OS irrespective of DBL maturity; this was particularly evident for nivolumab. This is an important finding given the recommendation that the exponential function should be considered the default parametric function for long-term survival extrapolations [33] and has been chosen as the preferred method in a number of IO HTA submissions [8, 9, 34, 35]. Note also that the cut-off point chosen for switching from the KM data to the exponential distribution can have a substantial impact on the survival estimates. However, exploratory post hoc analyses did not alter the conclusion that PieceW-exp consistently underestimated long-term survival.

In contrast to the 39-month DBL, the RBLM method failed to provide accurate estimates of survival for the 15- and 27-month DBLs. Performance on these earlier DBLs was similarly poor as for the PieceW-exp method. Internal validation and lifetime survival results showed that this could have been caused by the change in selection of functional form. In the 15- and 27-month DBLs, the curve with the lowest AIC for the non-responder group for both nivolumab and everolimus was an exponential curve. For the 39-month DBL, a log-logistic model performed better on both arms, which, according to the interval validation, provided more accurate estimates than the exponential models. This indicates that the hazards for the non-responder group were poorly represented by the exponential curve. Furthermore, heterogeneity among the non-responders, which consisted of those with stable disease, progressed disease and non-evaluable patients, may be a factor. Given these uncertainties, further research on the applicability of the RBLM method for IO therapies is warranted, particularly when follow-up is limited.

Although the MM method provided relatively accurate survival estimates for nivolumab based on the 39-month DBL, it significantly overestimated mean survival based on the 15- or 27-month DBLs. This may be due to the degrees of freedom provided by the MM method, which can lead to a potential overidentification of short-term trends in immature datasets even though inclusion of each parameter to be estimated is penalized in the definitions of AIC and BIC during the statistical fit. Future research could explore whether extrapolations using the MM method can be restrained when data maturity is low, for example by using external data. A quantitative approach with a Bayesian framework would avoid introducing subjectivity into the parametrization or selection of the models.

Within these analyses, the more complex methods—those with more degrees of freedom and a higher flexibility such as the RBLM and MM—did not have greater predictive accuracy than the SPM when fitted to the 15- and 27-month DBL. This may partly be explained by the PH assumption being maintained in this case study and the limited follow-up of these DBLs, as previously discussed. Further, it is important to be aware that more complex methods are typically associated with more uncertainty around their estimates.

Another consideration is the selection process for the best-fitting model. For the RBLM and especially the MM, many of the model parametrizations provided a virtually identical fit to the observed immature data. Absolute AIC and BIC differences among several of the best-fitting models were below one, implying none of the models significantly outperformed the others [36, 37]. Long-term survival estimates from these models were substantially different, though, and post hoc exploratory analyses showed that several of the models that were not selected did in fact provide close estimates of long-term survival. This suggests that thorough guidance on model selection for more complex methods is crucial and that current procedures may be improved.

These results are aligned with recent previous studies into survival extrapolation methods [4, 5, 38]. In a case study of ipilimumab in melanoma, Bullement et al. [4] found that complex flexible methods were more reflective of complex survival patterns but had poor predictive accuracy if not informed by external data. In this case study, similar challenges were found for the RBLM and MM methods.

A case study investigating survival of patients receiving durvalumab for pre-treated non-small-cell lung cancer found that RBLM, MM, and cure methods provided long-term survival estimates that were closer to observed OS rates in a later DBL than the SPM and spline methods [5]. However, the estimates from various RBLM, MM, and cure models differed considerably from each other, thus retaining uncertainty about future survival benefit. Similar variations between models, even within a single method, were also observed in our case study of pre-treated aRCC.

A review and validation study of 11 single technology appraisals of IO therapies demonstrated that, although the SPM and PieceW methods performed reasonably well at predicting long-term survival, the long-term survival estimates tended to be underestimated [38]. This is in line with our findings, where the SPM and PieceW methods consistently underestimated nivolumab survival, though at varying degrees.

A limitation of this case study is the limited external validation. Expert validation involved a single clinical expert, and the trials used for comparison used different doses of nivolumab. Involvement of multiple experts could increase the robustness of the survival estimates provided, though it is not clear how many experts would be required to obtain a robust estimate. To partially mitigate this limitation, the expert was asked to provide lower and upper limits to the survival estimates, a method derived from the Sheffield elicitation framework [27]. Furthermore, the expert opinion was only used for validation purposes and not for parametrization of the models, as Bayesian analyses were not within the scope of this study. Investigating the robustness of clinical expert opinion and evaluating formal frameworks for eliciting such input is an area of ongoing research [27, 39].

To ensure consistency of the analysis, external data were not used in determining the best performing distributions beyond a visual assessment of face validity. In practice, the plausibility of the survival extrapolation receives substantial scrutiny. An additional important point of future research would be the evaluation and validation of using external data, such as real-world data, to inform survival extrapolations. External data may be leveraged in multiple ways, ranging from inputs for model parametrization to model validation, but there is a paucity of evidence on which methods are robust and valid in the context of survival extrapolations in oncology. Given the potential value of external inputs such as real-world data, especially for the more complex survival extrapolation methods, further research is warranted.

5 Conclusion

This case study in pre-treated aRCC found that the accuracy of OS predictions for nivolumab differed considerably between extrapolation methods, whereas predictions for everolimus were generally accurate. Furthermore, all methods underestimated survival for nivolumab, except for MM. This adds to the growing evidence that suggests that the extrapolation method required for IO therapies may need different considerations than for therapies with other mechanisms of action, such as targeted therapies or chemotherapies.

Of special note is the poor performance of the PieceW-exp method, irrespective of the maturity of the dataset. Thus, use of the PieceW-exp method should be discouraged for extrapolating IO monotherapy, unless robust contrary evidence for the use of this approach is presented.

Methods providing more degrees of freedom in the extrapolation, such as RBLM and MM, provided accurate survival estimates based on the 39-month DBL but did not outperform the dependent log-logistic SPM in this case study. The dependent log-logistic SPM performed better than the RBLM and MM, based on the 15-month and 27-month DBLs. This may be due to the use of subsequent IO therapies, which could have biased survival in the everolimus arm. As data matured, the survival estimates from the RBLM and MM methods changed considerably, partly driven by a change in the selection of the best functional form. This highlights the need to include external evidence to inform the parametrization of these extrapolation methods that allow for more degrees of freedom.

Declarations

Funding This study was supported by Bristol-Myers Squibb Company.

Conflict of interest Sven L. Klijn, Elisabeth Fenwick and Sonja Kroep are employees of Pharmerit - an OPEN Health Company, which received payment from Bristol Myers Squibb to conduct the analyses. John Borrill, Kasper Johannesen, Christopher Kiff, Murat Kurt, and Bill Malcolm are employees and shareholders of Bristol Myers Squibb.

Data availability Bristol Myers Squibb Company policy on data sharing may be found at https://www.bms.com/researchers-and-partners/independent-research/data-sharing-request-process.html.

Author contributions All authors were involved in the conceptualization and design of the work. SLK and SK performed data analysis. All authors were involved in data interpretation. All authors reviewed and revised the manuscript.

Acknowledgements Jessica May and Kyna Gooden supported the data acquisition. Esra Cakar, Lisanne Verburg, and Poojee Sudhapalli contributed to the analyses of the study. Dr. Michael B. Atkins from Georgetown University provided clinical input for the external validation of the long-term survival extrapolations. Einar Torkilseng and Sébastien Branchoux contributed to the writing of this article.

References

1. 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. 2020;38(2):193–204.
2. Ferrara R, Pilotto S, Caccese M, Grizzi G, Sperduti I, Giannarelli D, et al. Do immune checkpoint inhibitors need new studies methodology? J Thorac Dis. 2018;10(Suppl 13):S1564–80.
3. Kaufman HL, Atkins MB, Subedi P, Wu J, Chambers J, Joseph Mattingly T 2nd, et al. The promise of immuno-oncology: implications for defining the value of cancer treatment. J Immunother Cancer. 2019;7(1):129.
4. Bullement A, Latimer NR, Bell GH. Survival extrapolation in cancer immunotherapy: a validation-based case study. Value Health. 2019;22(3):276–83.
5. Ouwens M, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. Pharmacoeconomics. 2019;37(9):1129–38.
6. Sculpher M, Palmer S. After 20 years of using economic evaluation, should NICE be considered a methods innovator? Pharmacoeconomics. 2020;38(3):247–57.
7. Latimer N. NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data. 2013.
8. Gallacher D, Auguste P, Connock M. How do pharmaceutical companies model survival of cancer patients? A review of NICE single technology appraisals in 2017. Int J Technol Assess Health Care. 2019;35(2):160–7.
9. Kroep S, Kiff C, Kraan C, Bianco M, Johannesen K, Kurt M, et al. PCN451. Modeling the survival benefit of immuno-oncologic therapy: a review of methods used in NICE single technology appraisals. Value Health. 2019;22:S523–4.
10. Latimer N. NICE DSU Technical support document 14: survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data. 2011.
11. Woods B, Sideris E, Palmer S, Latimer N, Soares M. NICE DSU Technical support document 19: partitioned survival analysis for decision modelling in health care: a critical review. 2017.
12. Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, et al. Nivolumab versus everolimus in advanced renal-cell carcinoma. N Engl J Med. 2015;373(19):1803–13.
13. National Institute for Health and Care Excellence. Final appraisal determination: nivolumab for previously treated advanced renal cell carcinoma. 2016.

14. Statens legemiddelverk. Hurtig metodevurdering: Nivolumab (Opdivo) til andrelinjebehandling av avansert nyrecellekarsinom - Vurdering av innsendt dokumentasjon. 2016.
15. Tandvårds- och Läkemedelsförmånsverket. Underlag foör beslut i landstingen: Opdivo som monoterapi är indicerat för behandling av vuxna med avancerad njurcellscancer efter tidigare behandling. 2016.
16. Motzer RJ, Tykodi SS, Escudier B, Oudard S, Hammers HJ, McDermott DF, et al. Final analysis of the CheckMate 025 trial comparing nivolumab (NIVO) versus everolimus (EVE) with > 5 years of follow-up in patients with advanced renal cell carcinoma (aRCC). J Clin Oncol. 2020;38(6_suppl):617.
17. Plimack ER, Motzer RJ, Escudier B, Sharma P, McDermott DF, George S, et al. Two-year efficacy and safety update: phase III CheckMate 025 study of nivolumab vs everolimus in patients with advanced renal cell carcinoma (aRCC). ESMO. 2016.
18. Sharma P, Tykodi SS, Escudier B, Carducci M, Oudard S, Hammers HJ, et al. Three-year efficacy and safety update from the phase III CheckMate 025 study of nivolumab versus everolimus in patients with advanced renal cell carcinoma. In: 16th International kidney cancer symposium. 2017.
19. 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. Stat Med. 2002;21(15):2175–97.
20. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–26.
21. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19(6):716–23.
22. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
23. Mazza C, Escudier B, Albiges L. Nivolumab in renal cell carcinoma: latest evidence and clinical potential. Ther Adv Med Oncol. 2017;9(3):171–81.
24. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med. 2012;366(26):2443–54.
25. Motzer RJ, Rini BI, McDermott DF, Redman BG, Kuzel TM, Harrison MR, et al. Nivolumab for metastatic renal cell carcinoma: results of a randomized phase II trial. J Clin Oncol. 2015;33(13):1430–7.
26. McDermott DF, Motzer RJ, Atkins MB, Plimack ER, Sznol M, George S, et al. Long-term overall survival (OS) with nivolumab in previously treated patients with advanced renal cell carcinoma (aRCC) from phase I and II studies. J Clin Oncol. 2016;34(15_suppl):4507.
27. O'Hagan A, Oakley JE. SHELF: the Sheffield Elicitation Framework; 2019. http://www.tonyohagan.co.uk/shelf/. Accessed Sept 2019.
28. National Office for Statistics. National life tables: UK 2019.
29. Bristol-Myers Squibb. Bristol-Myers Squibb receives FDA approval for Opdivo (nivolumab), the only treatment to deliver significant overall survival in advanced renal cell carcinoma vs. a standard of care, in patients who have received prior anti-angiogenic therapy. [2015 January 2020]. https://news.bms.com/press-release/bristol-myers-squibb-receives-fda-approval-opdivo-nivolumab-only-treatment-deliver-sig.
30. European Medicines Agency. Opdivo EPAR2020.
31. Motzer RJ, Tykodi SS, Escudier B, Carducci M, Oudard S, Hammers HJ, et al. Long-term follow-up of nivolumab versus everolimus in patients with advanced renal cell carcinoma: the phase 3 CheckMate 025 trial. In: 18th International Kidney Cancer Symposium. Miami, FL, USA2019. p. Poster.
32. Liverpool Reviews and Implementation Group. Atezolizumab for treating locally advanced or metastatic non-small cell lung cancer after chemotherapy [ID970]. 2017.
33. Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. Med Decis Mak. 2014;34(3):343–51.
34. National Institute for Health and Care Excellence. TA531: Pembrolizumab for untreated PD-L1-positive metastatic non-small-cell lung cancer. 2018.
35. National Institute for Health and Care Excellence. TA520: Atezolizumab for treating locally advanced or metastatic non-small-cell lung cancer after chemotherapy. 2018.
36. Burnham KP, Anderson DR. Multimodel inference. Sociol Methods Res. 2016;33(2):261–304.
37. Raftery AE. Bayesian model selection in social research. Sociol Methodol. 1995;25.
38. Bullement A, Meng Y, Cooper M, Lee D, Harding TL, O'Regan C, et al. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available? J Med Econ. 2019;22(3):205–14.
39. Cope S, Ayers D, Zhang J, Batt K, Jansen JP. 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. 2019;19(1):182.

Affiliations

Sven L. Klijn[1] · Elisabeth Fenwick[2] · Sonja Kroep[1] · Kasper Johannesen[3] · Bill Malcolm[4] · Murat Kurt[5] · Christopher Kiff[4] · John Borrill[4]

✉ Sven L. Klijn
sklijn@pharmerit.com

[1] Pharmerit - an OPEN Health Company, Marten Meesweg 107, 3068 AV Rotterdam, The Netherlands

[2] Pharmerit - an OPEN Health Company, Oxford, UK

[3] Bristol Myers Squibb, Stockholm, Sweden

[4] Bristol Myers Squibb, Uxbridge, Greater London, UK

[5] Bristol Myers Squibb, Princeton, NJ, USA