Medical Decision Making

Use of advanced flexible modelling approaches for survival extrapolation from early follow-up data in two nivolumab trials in advanced NSCLC with extended follow-up

Journal:	*Medical Decision Making*
Manuscript ID	MDM-21-499
Manuscript Type:	Original Research Article
APPLICATION AREAS:	PHARMACEUTICALS--PHARMACIST, OUTCOMES RESEARCH, ONCOLOGY
DETAILED METHODOLOGY:	ECONOMICS (HEALTH), Survival Analysis < STATISTICAL METHODS, Bayesian Statistical Methods < STATISTICAL METHODS

Use of advanced flexible modelling approaches for survival extrapolation from early follow-up data in two nivolumab trials in advanced NSCLC with extended follow-up

Objectives

Immuno-oncology (IO) therapies are often associated with delayed responses that are deep and durable, manifesting as long-term survival benefits in patients with metastatic cancer. Complex hazard functions associated with IO treatments may limit the accuracy of short- and long-term extrapolations from standard parametric models (SPMs). We evaluated the ability of advanced flexible parametric models (FPMs) to improve survival extrapolations relative to SPMs using data from two trials involving pre-treated advanced non-small cell lung cancer (NSCLC) patients.

Methods

Our analyses used consecutive database locks (DBLs) at 2-, 3-, and 5-years' minimum follow-up from trials evaluating nivolumab versus docetaxel in patients with pre-treated metastatic squamous (CheckMate-017) and non-squamous (CheckMate-057) NSCLC. For each DBL, SPMs, as well as three FPMs – landmark response models (LRMs), mixture cure models (MCMs), and Bayesian multi-parameter evidence syntheses (B-MPES) – were estimated on nivolumab overall survival (OS). Performance of each parametric model was assessed by comparison of restricted mean survival time (RMST) over 5 and 20-year horizons, and of survival probabilities at key milestones, with observed 5-year Kaplan-Meier estimates and results obtained from validated SPMs using the latest available data-cut. Selection of validated SPMs, which have previously been used in economic analyses submitted to regulators and reimbursement agencies and later published in peer reviewed journals, was informed by cancer registry data and by longer follow up from CheckMate-003.

Results

For the 2- and 3-year DBLs of both trials, all models tended to systematically underestimate 5-year OS. Predictions from non-validated SPMs fitted to the 2-year DBLs were highly unreliable, whereas extrapolations from FPMs were much more consistent between models fitted to successive DBLs. For CheckMate-017, where an apparent survival plateau emerges in the 3-year DBL, MCMs fitted to this data cut estimated 5-year OS most accurately (11.6% vs 12.3% observed) and gave long-term predictions most similar to those from the 5-year validated SPM (20-year RMST: 30.2 vs 30.5 months). For CheckMate-057, where there is no significant evidence of a survival plateau in the early DBLs, only B-MPES was able to accurately predict 5-year OS (14.1% vs 14.0% observed [3-year DBL]), although 5-year RMST was overestimated (19.8 vs 19.4 months [3-year DBL]). B-MPES consistently resulted in conservative 20-year OS estimates owing to an increasing hazard rate from intermediate timescales onwards.

Conclusions

We demonstrate that the use of FPMs for modelling OS in NSCLC patients from early follow-up data can yield accurate estimates for RMST observed with longer follow-up, and provide

similar long-term extrapolations to externally validated SPMs based on later data-cuts. B-MPES generated reasonable predictions even when fitted to the 2-year DBLs of the studies, whereas MCMs were more reliant on longer-
term data to estimate a plateau, and therefore performed better from 3 years. Generally, LRM extrapolations were less reliable than those from alternative FPMs and validated SPMs but remained superior to non-validated SPMs. Our work demonstrates the potential benefits of employing advanced parametric models that formally incorporate external data sources, such as B-MPES and MCMs, to allow for accurate evaluation of treatment clinical- and cost-effectiveness from limited trial data.

Words: 496

Highlights

- Flexible advanced parametric modelling methods can provide improved survival extrapolations for immuno-oncology cost-effectiveness in health technology assessments from early clinical trial data that better anticipate extended follow-up.

- Advantages include leveraging additional observable trial data, the systematic integration of external data, and more detailed modelling of underlying processes.

- Bayesian multi-parameter evidence synthesis performed particularly well, with well-matched external data.

- Mixture cure models also performed well but may require relatively longer follow-up to identify an emergent plateau, depending on the specific setting.

- Landmark response models offered marginal benefits in this scenario and may require greater numbers in each response group and/or increased follow-up to support improved extrapolation within each subgroup.

INTRODUCTION

Health technology assessment (HTA) bodies typically require estimates of long-term effectiveness of novel treatments to assess their value over patients' lifetimes. Reliable long-term extrapolations of patient survival are critical in securing support from agencies for new treatments. However, survival data available for HTAs are often limited, with relatively short follow-up available when treatments are being assessed for clinical- and cost-effectiveness by HTA bodies. While longer term follow-up may continue in trials once the primary endpoints have been met, there is potential benefit to improving survival extrapolations based on early follow-up data to provide patients with more timely access to improved therapies through earlier evidence of cost-effectiveness.

Standard parametric models (SPMs) are frequently used for survival extrapolations as outlined, for example, in UK National Institute for Health and Care Excellence (NICE) DSU TSD 14 [1]. In recent years there has also been increased use of more flexible models that address some of the limitations that may be encountered when applying SPMs and, in particular, can accommodate more complex hazard functions [2, 3]. The complexity of the hazard function may vary between indications and can also depend on the mechanism of action of a treatment. For instance, immuno-oncology (IO) therapies are typically associated with delayed responses that are deep and durable, and may therefore lead to long-term survival benefits in some patients with metastatic cancer, and hence result in complex patterns of survival [4]. A similar effect may occur when there is unexplained heterogeneity in the patient population, for example, with a subset of patients (that cannot be identified *a priori*) demonstrating long-term survival.

SPMs do not account for heterogeneity in patient response to treatment and so may struggle to accurately predict long-term survival. In addition to providing a poorer fit to observed data, SPMs may also underestimate the uncertainty in extrapolations, which may in turn result in an exaggerated impression of accuracy in health economic evaluations. Furthermore, standard models constrain the possible shape of hazard curves, which may provide an adequate fit to observed data but nonetheless lead to implausible extrapolations. Examples of alternative parametric models that are able to appropriately capture more complex hazard functions, such as those arising from heterogeneous response and long-term survivorship, include mixture cure, landmark response, and Bayesian models.

The application of more flexible parametric models to address the challenge of extrapolating survival for immuno-oncology (IO) therapies has received recent attention [4-6]. IO therapies are one example where heterogeneity is anticipated and has been observed, with a substantial fraction of metastatic patients attaining long-term survival in certain tumors [7-10]. Nivolumab, a programmed cell death 1 (PD-1) inhibitor that works by disrupting the PD- 1-mediated signaling between the PD-1 receptor (expressed on active T cells) and its ligands PD-L1 and PD-L2 (expressed on tumor cells), thereby restoring T-cell anti-tumor immunity, is an established IO therapy with proven efficacy in non-small cell lung cancer (NSCLC) patients [8].

CheckMate-003 was an open-label Phase I trial in subjects with advanced malignancies,

including NSCLC, treated with nivolumab. The results of this trial indicated that a subgroup of subjects exhibited a deep and durable response to treatment [11]. The later Phase III trials CheckMate-017 and CheckMate-057 echoed this pattern [12]. Figure 1 shows the evolving overall survival (OS) estimates from the three trials. Landmark analyses of these trials have shown significantly greater progression-free and overall survival for nivolumab patients compared to patients treated with taxane-based chemotherapy agents [8, 12]. SPMs typically fail to capture such plateaus in the survival probability, resulting in inaccurate estimates of long-term survival, and do not model the observed unexplained heterogeneity between the underlying patient groups. Due to the relative novelty of immunotherapies in NSCLC, there is value in comparing the performance of different advanced, flexible modelling approaches for modelling lifetime survival. The availability of successive follow-up from two trials, CheckMate-017 and CheckMate-057, each with a minimum follow-up of 5 years, one exhibiting apparent early plateauing and the other not, offered a unique opportunity to test the ability of innovative flexible approaches to improve upon the accuracy of the long-term projections from earlier databases, relative to SPMs.

The purpose of this work was to evaluate whether advanced flexible parametric models (FPMs) provide more accurate estimates of long-term survival for second-line NSCLC patients at earlier follow-up than SPMs. This paper focuses on the use of landmark response models (LRMs), mixture cure models (MCMs), and Bayesian multi-parameter evidence synthesis (B-MPES) as alternatives to SPMs when extrapolating survival data in NSCLC. LRMs and MCMs explicitly model heterogeneity by introducing latent subpopulations. MCMs and B-MPES directly utilize longer-term external data to supplement the trial data. All three methods are included in NICE guidance for flexible methods [2]. The principal datasets for our analysis are the successive database locks (DBLs) of CheckMate-017 and CheckMate-057 – two pivotal trials evaluating nivolumab versus docetaxel in patients with pre-treated metastatic squamous and non-squamous NSCLC, respectively – at 2-, 3- and 5-years' minimum follow-up.

METHODS

This study evaluated the performance of FPMs by applying selected methods to the initial OS data for the nivolumab arms of two clinical trials and comparing the results to longer follow-up data from these studies [12]. This section provides a description of the clinical trials that are of primary interest in this work, the follow-up that was available at the initial and later data-cuts for this pair of trials, the external data sources that were leveraged in the FPMs, and how the FPMs and SPMs were estimated using these data.

The approaches by which advanced parametric methods allow for improved modelling of trial data and representation of complex hazard functions can be broadly divided into two categories. Some methods, such as MCMs and B-MPES, formally incorporate external data via modelling assumptions and associated parameters that link the various sources. Other methods, including LRMs, do not use external data but instead leverage additional observable information from the study dataset of interest in a way that may allow for representing more complicated features of

the survival data, such as heterogeneity in patient response. Furthermore, external data are frequently used indirectly to inform model selection by supporting post-hoc validation of the survival probabilities estimated by parametric models. The use of external data is particularly useful when only limited trial follow-up data is available, as they can provide validation of the resulting extrapolations for survival probabilities and hazard rates from short-term data [13]. Such an approach was previously taken in relation to the CheckMate-017 and CheckMate-057 cost-effectiveness models used in HTA submissions, where external validation of parametric models was carried out using NSCLC data from the longer term Phase I trial CheckMate-003, as well as longer term data from the US National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program and the national cancer registries of Sweden and Norway [12, 14-18]. The nivolumab OS parametric models used in these cost-effectiveness models are herein referred to as 'validated SPMs', whereas those that do not leverage external data are referred to as 'non-validated SPMs' for the purpose of this research.

Clinical trials

CheckMate-017 (NCT01642004) and CheckMate-057 (NCT01673867) are randomized, open-label, international phase III studies comparing nivolumab vs. docetaxel in second-line squamous-cell and non-squamous-cell NSCLC, respectively [19, 20]. In the CheckMate-017 trial, 272 patients underwent 1:1 randomization to nivolumab at 3mg/kg every 2 weeks or docetaxel at 75mg/m^2 every 3 weeks. In the CheckMate-057 trial, 582 patients underwent 1:1 randomization to nivolumab at 3mg/kg every 2 weeks or docetaxel at 75mg/m^2 every 3 weeks.

The principal efficacy analyses in both trials were performed after approximately one-year minimum follow-up and annual updates were subsequently provided for the regulatory and HTA reviews. Patient follow-up beyond the primary analysis timepoint extended to a minimum follow-up of 5 years in both trials, with annual OS updates provided in a series of publications [8, 12].

Table 1 summarizes the characteristics of the 272 patients randomized in CheckMate-017 and the 582 patients randomized in CheckMate-057. Baseline age profiles were similar across both trials, but there were slight differences in stage (92% Stage IV in CheckMate-057 vs 80% in CheckMate-017), and some differences in sex (45% female in CheckMate-057 vs 24% in CheckMate-017). Kaplan-Meier (KM) estimates of OS in both trials are illustrated in Figure 2. There is some apparent volatility in OS estimates at the earliest DBLs. For example, the 1- and 2-year DBLs for CheckMate-057 generate low OS point estimates towards the end of the available follow-up, although the curve begins plateauing robustly with 3-year follow-up. Plateauing of the CheckMate-017 OS KM curve occurs relatively early compared to CheckMate-057, although 4-year OS probabilities for the two trials are very similar. Few patients were lost to follow-up in either trial, including in the 5-year DBL. From the combined dataset for the nivolumab treatment arms of the CheckMate-017 and CheckMate-057 studies, 17 (4%) of patients received subsequent immunotherapy and 185 (43%) of patients received subsequent chemotherapy [12]. More subsequent therapy options are available for non-squamous patients,

which may contribute to the improved OS in CheckMate-057 compared to CheckMate-017 at earlier times. However, it is anticipated that OS probabilities in second line NSCLC at intermediate and longer timescales are driven primarily by the proportion of responders to nivolumab, which could explain the observed greater conditional OS probabilities for CheckMate-017 compared to CheckMate-057 patients in the later DBLs.

Landmark response model (LRM)

When it is anticipated that a treatment, such as IO, may have a deep and durable response in a subset of the patient population, we may wish to leverage this information in order to explicitly account for response heterogeneity and hence improve survival extrapolations. Landmark response modelling allows for disaggregation of the study population according to an observed response classification at some "landmark" time point. Survival probability estimates *after* this time point are modelled separately for the responder and non-responder subpopulations. This avoids the "immortal time bias" inherent in assessing survival probabilities at earlier times according to a post-baseline characteristic (implicitly, subjects must have survived to this timepoint in order to be classified, thereby inferring nil hazard up to this time).

Estimating the landmark response models

LRMs have previously been applied to the pooled OS data from CheckMate-017, CheckMate-057, Phase I trial CheckMate-003, and Phase II Trial CheckMate-063 [12]. Here, a landmark time point of six months was used to allow sufficient time for most responders to be identified, while retaining sufficient follow-up to assess the impact of response classification on long-term survival. The selection of a landmark timepoint is a critical decision, and must support an assumption of homogeneity with respect to the endpoint up to this time, and heterogeneity thereafter. Moreover, since only survival data after the landmark time point and for the relevant subpopulation is utilized in the modelling of responders and non-responders, long-term survival probability estimates in LRMs can be associated with high uncertainty. Hence, it is desirable to select the earliest available time point at which response classification can be accurately assessed. For the present analysis, patients were classified as responders or non-responders according to whether they had a RECIST v1.1 classification of complete/partial response or stable disease/progressive disease, respectively, at six months' follow-up, based on earlier investigations.

Survival up to the six-month landmark point was estimated using the Kaplan-Meier method. Conditional survival from this landmark timepoint onwards was estimated using candidate SPMs, namely exponential, Weibull, gamma, log-logistic, log-normal, Gompertz, generalized gamma, and generalized F distributions. Model selection for the separate response groups was conducted on the basis of Akaike's Information Criterion (AIC), plausibility of hazard extrapolations and long-term OS estimates, and visual inspection of model residuals. Composite conditional survival estimates were obtained by weighting according to the response classification split at the six-month landmark.

Mixture cure model (MCM)

MCMs are a type of parametric survival model which, similar to LRMs, account for heterogeneity in response to treatment by introducing subpopulations with distinct survival probability functions. In an MCM, however, the underlying classification of patients into "cured" and "uncured" groups is latent and modelled at a population level, and the division of patients into these groups is assumed to hold at all times [21, 22]. The survival probability of the "cured" population is represented by a background mortality function that is parameterized in a separate initial step. While an outright cure may be unlikely in advanced cancer, such as metastatic NSCLC, MCMs can capture observed or anticipated plateaus in the survival curve and may still provide reasonable survival extrapolation over a limited time-horizon if prolonged response to treatment is anticipated. In this sense, it is arguably more accurate to use the term "durable responders" rather than "cured patients". The MCM survival probability function can be expressed as follows:

$$S(t;\pi, \boldsymbol{\theta}) = \pi S_c(t) + (1 - \pi)S_u(t;\boldsymbol{\theta}),$$

where π is the proportion of patients in the "cured" group (the "cure fraction"), S_c is the parametric survival function for the "cured" group (i.e., the background mortality), and S_u is the parametric survival function for the "uncured" group, with vector of parameters $\boldsymbol{\theta}$. After the background mortality has been specified, the cure fraction, π, and free parameters for the uncured population survival, $\boldsymbol{\theta}$, are estimated concomitantly.

Estimating the mixture cure models

We estimate the "cured" population parametric survival function using World Health Organization (WHO) life table data matched to the CheckMate-017 and CheckMate-057 datasets based on the age, sex, and country compositions of the trial populations. With the fixed background mortality thus specified, the cure fraction and parameters for the "uncured" population survival function were determined using maximum likelihood estimation. Additional detail on the estimation procedure for both the background mortality and the MCM survival function can be found in the supplementary materials. Each MCM was fit using one of five candidate SPMs recommended by NICE [1] for the "uncured" survival curve: the exponential, Weibull, log-logistic, log-normal, and Gompertz distributions. We refer to these models as the "exponential MCM", "Weibull MCM", and so forth. We did not consider the generalized gamma or generalized F distributions for the "uncured" survival curve, to balance between the risk of using an overly flexible uncured survival parameterization (which could yield estimates for the cure fraction that are associated with high variance) versus an overly rigid parameterization (which could be too restrictive to capture the hazard function for the uncured population). We select the best fitting of these MCMs considering the 2- and 3-year DBLs for OS in CheckMate-017 and CheckMate-057. Reliable estimation of the MCM requires that follow-up be sufficiently long for a plateau to be observable in the Kaplan-Meier curve [21]. For this reason, we also estimate MCMs on the progression-free survival (PFS) data for each DBL, to inform model selection under the assumption that PFS may be a leading indicator for OS.

Bayesian multi-parameter evidence synthesis (B-MPES)

B-MPES is an advanced parametric method that explicitly integrates patient-level data from clinical trials with external data and key clinical assumptions via a Bayesian framework. B-MPES was originally described by Guyot et al [23], who adapted the Royston and Parmar cubic spline model [24] to combine real world evidence in the form of cancer registry data with data from clinical trials to generate reliable long-term survival extrapolations for head and neck cancer patients. Vickers [25] included the B-MPES method in a benchmarking study to assess the performance of parametric methods that incorporate long-term external data in extrapolating simulated oncology survival data. The B-MPES method was found to perform favourably compared to alternative methods when relevant external information sources were incorporated appropriately and the treatment effect was relatively small.

The B-MPES method is based on fitting a suitably flexible parametric model, such as restricted cubic splines, to conditional survival probabilities that are obtained from a combination of clinical trial data and external data sources. Within specified time ranges, the predicted numbers of survivors in the control arm of the clinical trial are constrained to be equal to, or offset from, the numbers of survivors derived from an external data source (e.g., cancer registry or general population mortality data), which are sampled from a binomial distribution. The method also incorporates an explicit treatment waning effect via a tapering hazard ratio for the study treatment and control arms that converges to unity at a specified time, with precision expressed via a normal distribution and associated standard deviation hyperparameter. Prior distributions are also applied to the spline basis function coefficients, providing a further means to incorporate *a priori* knowledge and associated uncertainty, for example to reflect clinical expectation.

External data for B-MPES

Lung cancer relative conditional survival was estimated using data from the US National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) program, utilizing SEER*Stat statistical software [26]. These estimates are relative to general population mortality. Consistent with baseline study population characteristics, data were extracted for all available subjects aged 40 – 79 with a diagnosis of American Joint Committee on Cancer (AJCC) Stage IIIB or IV lung cancer of either squamous or non-squamous histology. These data were used to estimate one-year relative conditional survival probabilities (and associated variances) for each available year of follow-up.

General population mortality was estimated according to the distribution of country, sex, and age in each of the CheckMate-017 and CheckMate-057 trial populations separately using WHO life table data for 2013, corresponding to the commencement of both clinical trials. The WHO data is grouped into age bands of under 1 year, 1 to 4 years old, then five-year bands up to age 85, and finally aged 85 years and over. Conditional survival for each age band was modelled using a separate exponential model from which one-year conditional survival was estimated. One-year conditional survival for those aged 85 years and older, up to age 110 years, was estimated by geometric extrapolation of the exponential hazard rate.

Estimating the B-MPES models

B-MPES was performed using three main model constraints: patients in the docetaxel treatment arm were assumed to have one-year conditional survival probabilities that converged to those observed in the SEER lung cancer population after six years; the same docetaxel patients were assumed to have one-year conditional survival probabilities that were consistent with those of a matched general population sample after thirty years; and the treatment effect of nivolumab over docetaxel was assumed to have completely waned after six years – i.e., the hazard ratio is assumed to converge to 1 by this timepoint. This conservative treatment waning assumption was used as per Guyot et al [23], based on the maximum follow-up within the trials. Unlike for the other parametric methods considered in this work, docetaxel survival was modelled in order to provide linkage to the external data through the model structure.

Restricted cubic spline models with two internal knots were employed to model the relationship between cumulative hazards and time on a log scale, with independent spline models used for the separate treatment arms. Unlike Guyot et al, we did not constrain the treatment arms to have a common intercept, to allow a more flexible fit to the trial data and more valid comparison of the B-MPES method with the independent formulations of the other FPMs considered in this work. Boundary knots were placed at the earliest death in the trial and at 40 years, and internal knots were placed at half the maximum trial follow-up and at 18 years – mid-way between the extremes of the external data constraints. Weakly informative priors – specifically, diffuse normal distributions - were specified for the basis function coefficients of the splines.

The model was estimated in R using the Stan Bayesian modelling language via the Rstan package [27,28]. This program implements the No-U-Turn Sampler (NUTS) variant of the Hamiltonian Monte Carlo (HMC) algorithm [29]. Three Markov chains were used, each with 10,000 iterations and a 5,000-iteration warm-up. To ensure that valid starting values were used for spline parameters, these were initialized using maximum likelihood estimates from separate parametric bootstraps per chain. Model convergence was assessed by visual inspection of trace plots, comparison of within- and between-chain estimates using the $\hat{R}$ metric with a target value of one, estimation of effective sample size (ESS), and presence of any HMC divergent transitions. Definitions of these metrics and details of their assessment in this study are available in the Supplementary Materials.

Model comparisons

The performance of each FPM (LRM, MCM, and B-MPES), as well as the performance of SPMs, was assessed principally by comparison of restricted mean survival times (RMSTs) [30-32] obtained for models fitted to each of the 2- and 3-year DBLs with predictions from the validated SPM estimated using the 5-year DBL. Specifically, we examine model predictions for RMSTs estimated using 5- and 20-year time horizons. The 5-year RMST is a direct observable, since the model predictions can be compared with the KM estimate obtained from the 5-year DBL of the relevant trial. The 20-year RMST is highly relevant in health technology assessments, where economic models predict the cost-effectiveness of an intervention over a lifetime scale,

and the results can be strongly sensitive to the extrapolated survival probabilities. Moreover, the alternative assumptions of the various FPMs entail greatly differing approaches to modelling long-term survivorship, and SPMs frequently yield unreliable extrapolations. In addition, survival probability estimates were compared at 5, 10, 15, and 20 years, and extrapolated hazard rate estimates were inspected. All model comparisons relate to the nivolumab treatment arm only.

Comparison SPMs include: 1) non-validated SPMs selected based on AIC, assessment of proportional hazards, and visual inspection only; 2) HTA body models (ERG SPM) that were proposed by the NICE Evidence Review Group (ERG) [15]; and 3) validated SPM models selected by Bristol Myers Squibb (BMS) after comparisons with other trial data and external data sources [14-18]. In this previous work, dependent and independent parametric models were considered by BMS in model selection, although the focus of this paper is restricted to the nivolumab arm. Validated SPMs were those used in cost-effectiveness models submitted to NICE and other HTA bodies and later published in peer-reviewed journals, and were selected conservatively to avoid criticism by the ERG and other HTA reviewers [14-18].

RESULTS

Model fitting

Standard parametric models

SPMs selected for comparison are presented in Table 2. Independent models were fitted separately to each treatment arm, and dependent models included treatment effect as a covariate on the location parameter, although in the present work we are concerned only with the estimated SPMs for the nivolumab arm. ERG SPMs for the 2- and 3-year data cuts were recreated by applying the model assumptions proposed by the NICE ERG during nivolumab assessments for second-line treatment in advanced NSCLC, which pertained to 24- and 18-month DBLs for CheckMate-017 and -057, respectively [14-18]. In these appraisals, the ERG proposed a hybrid exponential model combining KM estimates up to 40 weeks in CheckMate-017 and to eight months in CheckMate-057 with an exponential distribution thereafter. The proportional hazards (PH) adjusted log-logistic model used as validated SPM for the CheckMate-017 3-year DBL was determined using a parametric log-logistic model fit to the docetaxel arm and a constant hazard ratio, estimated by Cox regression, applied to obtain nivolumab estimates.

Landmark response models

In both CheckMate-017 and CheckMate-057, 27% of nivolumab arm patients still being followed up at six months were classified as responders according to RECIST v1.1. The hazard ratio for responders vs non-responders within the nivolumab arm was highly significant for both trials, with values of 0.19 and 0.27 for CheckMate-017 and CheckMate-057, respectively. For the CheckMate-017 LRM, responders were modelled using an exponential model and non-responders using a log-normal model, for both the 2- and 3-year DBLs. For the CheckMate-057 LRM, both responders and non-responders were modelled using an exponential distribution for

the 2-year DBL. For the 3-year DBL, responders were modelled using a log-normal distribution and non-responders were modelled using an exponential distribution. In all cases, the selected models correspond to those with the lowest AIC, except for: CheckMate-017 responders for the 3-year DBL, where the model with second lowest AIC was selected, as the lowest AIC distribution (Gompertz) corresponded to a markedly high prediction for the 20-year RMST; CheckMate-057 non-responders (2-DBL), where the model with the second lowest AIC was selected, due to the hazard rate for the lowest AIC distribution (Gompertz) increasing rapidly to implausible levels; CheckMate-057 non-responders (3-year DBL), for which the lowest AIC distribution (generalized F) did not generate smooth predictions owing to numerical instability; CheckMate-057 responders (2-year DBL), where the residuals (deviances) for the exponential model showed the greatest consistency. Further details on LRM fitting, including AIC tables, are available in the supplementary material.

Mixture cure models

In the estimation of the MCMs for the CheckMate-017 and CheckMate-057 studies, the Gompertz distribution was selected to represent the survival in the "cured" population for both trials, based on lowest AIC score in the least-squares fit to the trial-matched WHO life table datasets. For the 2- and 3-year DBLs of CheckMate-017, considering the various candidate SPMs for survival in the "uncured" population, the lowest AIC MCM employed a log-logistic distribution. The resulting model is referred to herein as the "log-logistic MCM". For CheckMate-057, the lowest AIC model was the log-normal MCM for both the 2- and 3-year database locks. Since the improvement in AIC relative to other MCM parameterizations was quite modest for both CheckMate-017 and CheckMate-057, we also considered the fit of each of the MCMs for progression-free survival (PFS) as a potential leading indicator for OS, due to the earlier emergence of a KM plateau for the former outcome (see supplementary materials). We found that, for PFS, the log-logistic and log-normal MCMs were strongly preferred to alternative candidate MCMs in terms of AIC, for all DBLs of CheckMate-017 and CheckMate-057. This finding lends further support to our choice of MCMs. Due to the long tails of log-logistic and log-normal survival functions, these distributions allow for the possibility that some uncured patients may still have prolonged survival and declining hazards. As a result, these models may tend towards more conservative point estimates for the cure fraction, which is indeed the case in the present analysis. Further details on the procedures for fitting the MCMs are included in the supplementary materials.

Bayesian multi-parameter evidence synthesis

In the B-MPES model specification, trial-matched general population 1-year conditional survival probability estimates at 30 years, the timepoint at which the control arm estimates were constrained to match the extracted WHO life table data, were 0.730 and 0.753 for CheckMate-017 and CheckMate-057, respectively. SEER*Stat survival estimates were available up to 19 years follow-up. The 1-year SEER*Stat conditional survival probability estimates were 0.840 and 0.846 at 6 years (the initial timepoint at which the control arm estimates were constrained to match the extracted registry data), 0.863 and 0.870 at 10 years, and 0.869 and 0.879 at 15 years, for CheckMate-017 and CheckMate-057, respectively. The set of independent HMC trajectories used

to estimate the parameters of the B-MPES models showed good evidence of convergence. $\hat{R}$ estimates for the spline model parameters ranged between 1.000 and 1.002; trace plots showed no apparent trend, exhibited no obvious autocorrelation, and indicated good mixing across chains; ESS estimates were all large (all exceeded 3,800); and there were no divergent transitions. Further details on the procedures for fitting the BMPES model are included in the supplementary materials.

Comparison of models

CheckMate-017: short-term extrapolations

Predicted overall survival probabilities, hazard rates, and RMSTs from parametric models fitted to CheckMate-017 data are shown in Figure 3 and Figure 4 for the 2- and 3-year DBLs, respectively. For the 2-year DBL, extrapolation over a relatively short time horizon reveals significant discrepancies between the various approaches, with all models systematically underestimating the observed (i.e., Kaplan-Meier) 5-year OS probability (12.3%) and, excepting B-MPES, RMST (17.5 months). Notably, each of the FPMs show much better agreement with the (then yet to be observed) data from the 5-year DBL than any of the SPMs applied to the same dataset. The overall best-performing model is that based on B-MPES, which correctly predicts the RMST observed in the 5-year data (17.5 months), and also yields the most accurate prediction for 5-year OS probability (11.0%). In comparison, the LRM and MCM underestimate the 5-year OS probability (9.2% and 9.4%, respectively) and RMST (17.0 and 16.9 months, respectively) by a greater margin. Nonetheless, these FPMs yield markedly more realistic predictions than the non-validated and validated SPMs, which each predict a 5-year OS probability of 6.1% and RMST of 15.8 months. In contrast, the externally validated SPM applied to the 5-year DBL overestimates the 5-year OS probability (12.9%) and RMST (17.9 months).

For the 3-year DBL (Figure 4), the results of short-term extrapolations from the various models are more consistent, with a comparatively narrow range of estimates for 5-year RMSTs (17.4 months [validated SPM/B-MPES] – 17.7 months [LRM]). As for the 2-year DBL, all models provide a conservative estimate for the 5-year OS probability. The most conservative estimates arise from the ERG SPM (6.2%), validated SPM (10.2%), and from B-MPES (10.5%), and the most accurate predictions are now from the MCM (11.6%) and non-validated SPM (11.5%).

CheckMate-017: long-term extrapolations

Further differences between the alternative parametric models are apparent upon extrapolating predictions over a longer (namely, 20-year) time horizon, where the validated SPM fitted to the 5-year DBL estimates a 20-year OS probability of 4.1% and RMST of 30.5 months. Except for B-MPES, the predictions for these quantities differ significantly between models fitted to the 2- and 3-year DBLs, suggesting that, as is anticipated, 2-year minimum follow-up is insufficient to generate reliable estimates for lifetime survival in this indication. Hence, we focus on models fitted to the 3-year DBL (Figure 4). The estimates from the B-MPES approach are least sensitive to the inclusion of the additional data from a further one year of minimum follow-up (20-year OS

probability: 0.9% vs 0.9%; RMST: 25.0 vs 24.6 months [2- vs 3-year DBL]), since this method explicitly incorporates external data to inform the long-term predictions.

Of the parametric models fitted to the 3-year DBL, excluding the ERG SPM (which erroneously predicts 0.0% OS even after 15 years), the most conservative estimates for long-term survival arise from the B-MPES and LRM approaches, which predict 20-year OS probabilities of 0.9% and 1.1%, and 20-year RMSTs of 24.6 and 24.9 months, respectively. The MCM (20-year OS probability: 4.3%; RMST: 30.2 months) is a notable outlier, although this model most closely matches the validated SPM fitted to the 5-year DBL. The non-validated and validated SPMs yield intermediate estimates (20-year OS probabilities of 2.2% and 2.6%, respectively).

CheckMate-057: short-term extrapolations

The results from parametric models fitted to the CheckMate-057 data are illustrated in Figure 5 and Figure 6 for the 2- and 3-year DBLs, respectively. Among the models fitted to the 2-year DBL, the B-MPES estimates are in closest agreement with the 5-year KM data (OS probability: 14.7% vs 14.0%; RMST: 20.0 vs 19.4 months). Unlike B-MPES, the alternative parametric models fitted at 2-years minimum follow-up each underestimate OS when extrapolation is performed over a short time horizon. The next most accurate predictions are from the MCM and validated SPM (5-year OS probability: 10.3%; RMST: 18.7 mo), which are identical in this case, since the MCM estimates a zero cure fraction and the parametric form for the uncured population is the same as used in the validated SPM. The non-validated SPM performs very poorly, being unrealistically pessimistic (5-year OS probability: 3.5%; RMST: 17.3 mo), and the same is true for the ERG SPM (5-year OS probability: 4.9%; RMST: 17.6 mo).

For the 3-year DBL (Figure 6), the most accurate predictions over a 5-year time horizon are again those from B-MPES. With the additional year of minimum follow-up available, the B-MPES estimates have converged to very close agreement with the observed (i.e., KM) results for the 5-year DBL (OS probability: 14.1% vs 14.0%; RMST: 19.8 vs 19.4 months). The MCM now estimates a small but non-zero cure fraction, although 5-year predictions from the MCM and validated SPM remain consistent (OS probability: 11.8%; RMST: 19.4 mo), and these models are the next most accurate with respect to shorter-term extrapolations. Estimates from the non-validated SPM have changed most dramatically between the successive DBLs, which now yields a much more reasonable, though still pessimistic, prediction (5-year OS probability: 11.3%; RMST: 19.4 mo), while the ERG SPM remains poor.

CheckMate-057: long-term extrapolations

Similar to CheckMate-017, the long-term extrapolations of the candidate parametric models fitted to CheckMate-057 data are not stable between successive DBLs, with the exception of the B-MPES method (20-year OS probability: 1.4% vs 1.4%; RMST: 30.4 vs 29.9 months [2- vs 3-year DBL]). Hence, we focus on long-term extrapolations from models fitted to the 3-year DBL of CheckMate-057 (Figure 6). As for CheckMate-017, the long-term projections of the ERG SPM are spuriously low (20-year OS probability: 0.0%). The least conservative estimate for the 20-year RMST is that obtained from B-MPES (29.9 months), which also exceeds the estimate

corresponding to the validated SPM fitted to the 5-year DBL (28.6 months). However, it should be noted that the relatively high estimate for 20-year RMST from B-MPES is a result of higher predicted OS probability at short and intermediate times (i.e., less than around 10 years), including at the 5-year timepoint, where, unlike other models, B-MPES is in close agreement with the Kaplan-Meier result. At the 20-year timepoint, B-MPES in fact yields the most conservative estimate for OS (1.4%, compared to 1.8% for the 5-year validated SPM). In contrast, the non-validated and validated SPMs, as well as the LRM and MCM, underestimate the observed OS probability at the 5-year timepoint, but yield predictions that are less conservative than (or, in the case of the MCM, consistent with) the 5-year validated SPM at the 20-year timepoint. Moreover, each of the FPMs provide an estimate for 20-year OS probability that is more conservative, and hence more consistent with the 5-year validated SPM, than the non-validated and validated 3-year SPMs. This is true even for the MCM, where the small cure fraction and use of a Gompertz distribution for the cured population help to prevent overestimation of the proportion of long-term survivors.

DISCUSSION

Selection of the most appropriate OS extrapolation method for HTAs is often based on trial data with limited follow-up. The mechanism of action of immuno-oncology (IO) therapies may lead to complex OS hazard functions arising from treatment effects such as heterogeneity in long-term response, which may not be reflected in short-term follow-up and which SPMs may be insufficiently flexible to capture. FPMs incorporate external information or additional observable information from the dataset of interest, as well as further parameters that correspond to clinical assumptions, in a way that may allow for a more valid representation of complex hazard functions. We assessed the performance of three FPMs (LRMs, MCMs, and B-MPES), as well as non-validated and externally validated SPMs, fitted to 2- and 3-year DBLs of two similar trials for nivolumab in NSCLC that display somewhat different survival patterns at early times. In particular, we aimed to identify the earliest DBL at which plausible short- (5-year) and long- (20-year) -term extrapolations could be achieved with the use of advanced parametric methods, and thereby provide recommendations on the use of these methods in HTA assessments of immuno-oncology drugs.

Overall, we found that FPMs fitted to early DBLs were able to provide accurate predictions for the 5-year RMST that was observed with longer follow-up, for both studies considered. Whereas predictions from SPMs were highly variable between distributions fitted to the 2- and 3-year DBLs, dramatically underestimating 5-year OS probability in the former case, results for short-term extrapolations from FPMs were much more consistent between the successive DBLs. Notably, B-MPES was able to accurately predict 5-year RMST as early as the 2-year DBL in both trials. All parametric models tended to yield conservative estimates for 5-year OS probabilities and RMSTs.

The FPMs also provided long-term OS estimates (20-year RMST and OS probability estimates) on the 2- or 3-year DBLs that were in line with those produced by the validated SPM fitted to the 5-year DBLs of the respective trials. The validated and non-validated SPMs generally

underestimated long-term OS when applied to the 2-year DBLs, but gave significantly more optimistic lifetime projections when fitted to the 3-year DBLs. Overall, the FPMs more consistently provided better long-term extrapolations across all the datasets considered, although in some instances SPMs could perform well, especially when applied to more mature data-cuts.

LRM predictions for both short- and long-term quantities were less pessimistic, and thus more realistic, when utilizing the more mature (i.e., 3-year, rather than 2-year) DBLs. LRMs are susceptible to generating unreliable extrapolations when fitted to immature datasets, since the problem of limited observations is compounded by the data loss from fitting parametric models for subpopulations after a landmark timepoint. In the case of the 2-year DBL with a 6-month landmark timepoint, 25% of the follow-up is not available for model estimation. In addition, separate parametric models are fitted for each response group, and responder groups frequently make up a minority of the patient population. When there are few events, as is liable to be the case for OS data with limited follow-up, the lack of information can lead to selection of simpler models and high uncertainty in resulting estimates. In this study, we found LRMs to perform better in CheckMate-017 than in CheckMate-057; this may be anticipated given that the estimated hazard ratios when comparing responder and non-responder groups indicate greater heterogeneity between groups in CheckMate-017 than in CheckMate-057 (see supplementary material). Alternative response groupings according to trial characteristics may be associated with improved results if they maximize within-group homogeneity and between-group heterogeneity. Regardless, we would not recommend LRMs in this setting, as other FPMs provided more accurate predictions for 5-year RMST, and generated 20-year predictions that were in better agreement with the validated SPM fitted to the 5-year DBLs, for both trials and at all levels of follow-up considered.

The 5-year extrapolations from the MCM were the least conservative, and most accurate, of the candidate parametric models fitted to the 3-year DBL of CheckMate-017, which exhibits an apparent early plateau in the OS probability. Furthermore, extrapolation over a 20-year time horizon shows close agreement between the MCM and the 5-year validated SPM. This success is attributable to the MCM being able to represent the plateau naturally through a finite cure fraction, estimates for which appear to stabilize around the timescale of the 3-year DBL (see supplementary material, Figure 4). However, for the 2-year DBL, where a robust curative plateau has not yet emerged in the KM curve, the MCM predicts only a very small cure fraction, and hence a more realistic model is instead given by the B-MPES method. This problem of both short- and long-term predictions being highly sensitive to the value of the cure fraction limits the usefulness of applying MCMs to early follow-up data [13]. The MCM does not perform as well when applied to the CheckMate-057 dataset, for which the Kaplan-Meier plateau is slightly less pronounced, and therefore the curative hypothesis less appropriate, at early DBLs compared with CheckMate-017. Nonetheless, the 3-year MCM applied to CheckMate-057 yields long-term predictions that are highly consistent with the other FPMs and with the 5-year validated SPM, and the estimate for 5-year RMST matches observation, although the 5-year OS probability is significantly underestimated. This inaccuracy highlights an additional limitation of MCMs, namely that the dependability of their long-term survival extrapolations relies on the assumption that a

proportion of patients will be “cured” in the sense of achieving sufficiently durable response. However, it is difficult to assess the plausibility of this assumption. One option when assessing the plausibility of a durable response to treatment is to consider the use of surrogate endpoints such as PFS, as in the present work. Another approach is to explore parametric mixture models that do not incorporate a pre-specified background mortality, but instead the models for both subpopulations are estimated using the trial data. However, such models typically require a large amount of observations to reduce uncertainty and thus yield meaningful results.

We found a high level of consistency in the projections arising from B-MPES models fitted to successive DBLs, which are therefore able to generate reasonable short- and long-term extrapolations even when fitted to the 2-year DBLs. This stability is presumed to stem from the leveraging of external data to inform the extrapolation, in accordance with recommendations described in NICE TSD 21 [2]. In fact, the B-MPES approach yielded the least conservative and most realistic 5-year predictions for both 2- and 3-year data cuts of CheckMate-057, which does not exhibit early plateauing in OS. Although the B-MPES approach performs favourably in the present work, it is important to note that the method is highly reliant on a number of key elements, namely: availability and selection of appropriate external data; suitable and justifiable clinical assumptions (e.g., relating to treatment waning); and appropriate specification of the parametric survival function. The latter consideration is further complicated when a restricted spline function is applied, which is usual since the model then has adequate flexibility to fit to disparate data sources, because the number and placement of knots can appreciably influence results. Although external information sources are valuable for informing model predictions, it is nonetheless important to be aware of limitations in any datasets employed. In particular, available external data may reflect older treatments and standards of care, thereby resulting in conservative survival estimates. Indeed, for both trials analysed herein, the B-MPES approach results in the most conservative estimates for long-term survival out of all parametric models fitted to the 3-year DBLs (excepting the ERG SPM). This finding is attributable to the fact that only the B-MPES (and, for CheckMate-017, MCM) approach demonstrates the arguably expected qualitative trend that hazard rates increase slightly beginning at intermediate timescales (around 10 years), owing to age-related mortality.

In this study, we generally find that advanced parametric methods that directly leverage external data are able to generate more accurate extrapolations from less mature OS data. This contrasts with the ‘informal’ use of external data, for example by visual inspection, to inform SPM selection, which still led to systematic underestimation of short- and long-term OS when applied to the 2-year data-cut. While FPMs are therefore a very promising approach to perform short- and long-term survival extrapolations from early clinical trial data, these advanced methods involve numerous implementation decisions that are crucial to the quality of the resulting predictions. In choosing a FPM, care must be taken to verify that the clinical assumptions implicit in the proposed method are plausible. The external data that are integrated into the FPM must be relevant to the study population, otherwise the model extrapolations are liable to be poor. Although beyond the scope of the present work, it is also important to perform scenario and sensitivity analyses to ensure that the predictions resulting from FPMs are robust to the precise model specification. These checks should include variation of parameter (and, for Bayesian

methods, hyperparameter) values, as well as exploring alternative choices for underlying parametric forms of model components (including prior distributions for Bayesian methods). Similarly, it is instructive to assess the uncertainty associated with model outputs, which can be large when models have an excessive number of parameters, especially if a parameter corresponds to a clinical assumption that is invalid. Key assumptions, requirements, advantages, and disadvantages of the FPMs considered in the present work are summarized in Table 3.

Interesting opportunities exist to further develop the advanced parametric methods investigated in the present work, and to potentially hybridize methodologies by blending elements of different techniques. The FPMs could leverage treatment-specific external data, such as the Phase I and Phase II trial data that were utilized in selecting validated SPMs for the subsequent Phase III trials. External data could be used to mitigate the challenges in estimating survival for smaller responder groups in LRMs, while response classification could be used as a covariate in MCMs, alongside other factors, for example in estimating the cure fraction. MCMs could employ a Bayesian framework to inform the prior beliefs, according to expert opinion or other information sources, in connection to the cure fraction; a quantity that is challenging to estimate and strongly influences the model predictions. Furthermore, the application of external data could be more nuanced, for example by including a mortality adjustment in MCMs to allow survival in the "cured" population to differ from general population mortality, or by allowing for a 'random effect' in the pooling of data from various sources in B-MPES, which could be further extended to include multiple registries and/or clinical trials. There is also scope to modify some of the clinical assumptions that are implicit in the B-MPES approach. For instance, the treatment waning assumption could be relaxed so that the survival function for the immunotherapy arm is offset from, rather than matched to, that for the control arm, to avoid overly conservative long-term predictions.

Conclusions

This study demonstrates how FPMs, when employed and parameterized appropriately, can provide reliable short- and long-term OS extrapolations from trial data with limited follow-up. Whereas predictions from SPMs are frequently slow to converge with subsequent follow-up, FPMs can yield stable predictions even with relatively few observations, by supplementing immature trial data with relevant external data. Moreover, the additional flexibility of advanced parametric methods allows for representation of complex treatment effects such as heterogeneous response and long-term survivorship. Hence, use of FPMs may allow for accurate assessment of clinical- and cost-effectiveness using early trial data cuts, potentially resulting in expedited patient access to novel therapies. While it is important to note that the current analysis is limited to two clinical trials of nivolumab as a second line treatment in squamous and non-squamous metastatic NSCLC patients, it is reasonable to suggest that, considering a similar mechanism of action, results may hold for other immunotherapy indications.

Word count: 8229 (incl abstract and highlights, not incl tables or captions)

Table 1: Demographic and clinical characteristics of CheckMate-017 and CheckMate-057 patients

		CheckMate-017 (Squamous)		CheckMate-057 (Non-squamous)	
Characteristic	**Value**	**Nivolumab**	**Docetaxel**	**Nivolumab**	**Docetaxel**
N		135 (50%)	137 (50%)	292 (50%)	290 (50%)
Stage	Stage IIIB	29 (21%)	24 (18%)	20 (7%)	24 (8%)
	Stage IV	105 (78%)	112 (82%)	272 (93%)	266 (92%)
	Not reported	1 (<1%)	1 (<1%)	-	-
Sex	Female	24 (18%)	40 (29%)	141 (48%)	122 (42%)
	Male	111 (82%)	97 (71%)	151 (52%)	168 (58%)
Age	<65 years	79 (59%)	73 (53%)	184 (63%)	155 (53%)
	65-74 years	45 (33%)	46 (34%)	88 (30%)	112 (39%)
	75+ years	11 (8%)	18 (13%)	20 (7%)	23 (8%)

Table 2: Comparator standard parametric models for overall survival (nivolumab)

Dataset	Non-validated SPM	Validated SPM	ERG SPM
CheckMate-017			
2yr DBL	Dependent log-logistic	Dependent log-logistic	Independent hybrid exponential (40-week breakpoint)
3yr DBL	Dependent 2-knot hazard spline	PH-adjusted log-logistic	Independent hybrid exponential (40-week breakpoint)
5yr DBL	Dependent 2-knot hazard spline		
CheckMate-057			
2yr DBL	Independent 2-knot hazard spline	Independent log-normal	Independent hybrid exponential (eight-month breakpoint)
3yr DBL	Independent 3-knot odds spline	Independent log-logistic	Independent hybrid exponential (eight-month breakpoint)
5yr DBL	Independent 3-knot hazard spline	Independent log-normal	

Table 3: Summary table of the key features of the flexible parametric survival methods considered in this work.

Flexible parametric method	Supplemental data sources required	Key assumptions	Advantages	Limitations
Landmark response model (LRM)	- Additional trial data for patient characteristics at landmark timepoint, allowing for responder/non-responder classification	- Responder and non-responder groups are distinguishable and experience distinct hazard rates - Survival probability modelled as an aggregate for the whole study population before the landmark timepoint, and separately by response status thereafter	- response status of individual patients explicitly identified and used to improve modeling	- uncertainty liable to be high (data loss from landmark timepoint and splitting into subpopulations) - sensitive to choice of landmark timepoint - sensitive to choice of subpopulation aggregating
Mixture cure model (MCM)	- General population mortality (life table) data - Adjustments to general population data (optional)	- Curative effect (finite proportion of patients can be identified as experiencing background risk only)	- cure fraction parameter is intuitive and readily interpretable	- sensitive to value of cure fraction - can be sensitive to choice of parametric model for uncured population - requires sufficient data (observation of apparent plateau) to reliably estimate cure fraction
Bayesian multi-parameter evidence synthesis (B-MPES)	- Trial data for control arm - Registry mortality data corresponding to control arm for relevant indication - General population mortality data	- Treatment waning (converging hazard ratio for experimental vs control arm) - Trial data matched to registry and/or general population mortality data within specified time windows	- often outperforms alternative parametric models when applied to early data cuts - predictions less sensitive to successive follow-up than alternative methods (especially at intermediate and long timescales, where results are driven by external data)	- typically yields conservative long-term estimates (since these are based on control arm) - complicated to implement and analyse (many model components) - -sensitive to choice of timepoints where trial data is matched to external data - may be sensitive to number and location of knots in spline function - may be sensitive to choice of prior distributions

Figure 1: Kaplan-Meier overall survival estimates from the Phase I CheckMate-003, Phase III CheckMate-017, and Phase III CheckMate-057 studies. (A) CheckMate-003 (nivolumab arm), 5-year minimum follow-up. (B) CheckMate-017, 2-year minimum follow-up, demonstrates apparent early plateauing. (C) CheckMate-057, 2-year minimum follow-up, does not demonstrate plateauing in the nivolumab arm. (D) Nivolumab arms for: CheckMate-003, 6-year minimum follow-up; CheckMate-017, 5-year minimum follow-up; CheckMate-057, 5-year minimum follow-up. At these later database locks, all three trials demonstrate apparent plateauing in the nivolumab arm.

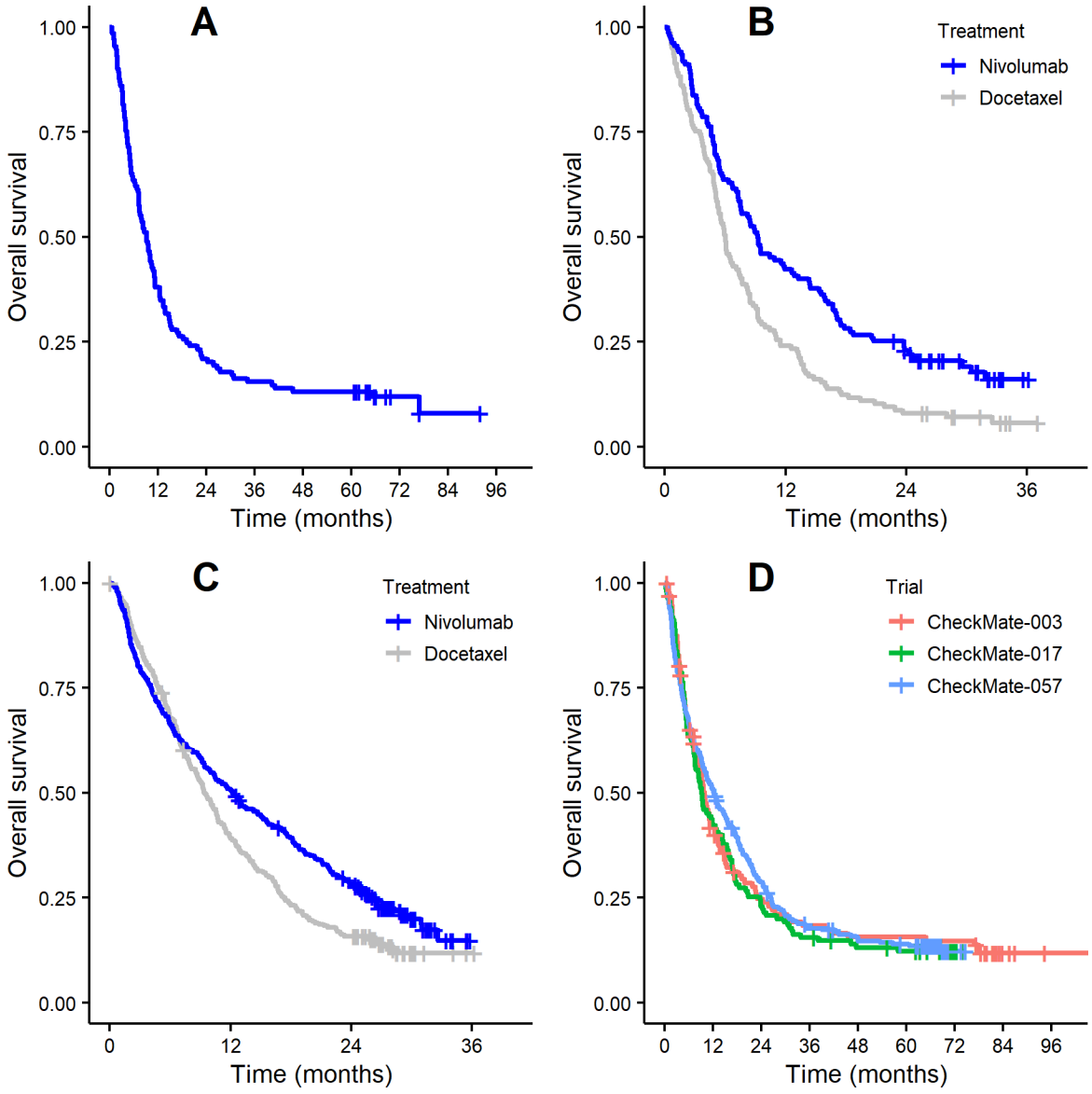


Figure 2: Kaplan-Meier overall survival estimates for the nivolumab treatment arms in the CheckMate-017 and CheckMate-057 studies, colored by successive database lock (DBL)

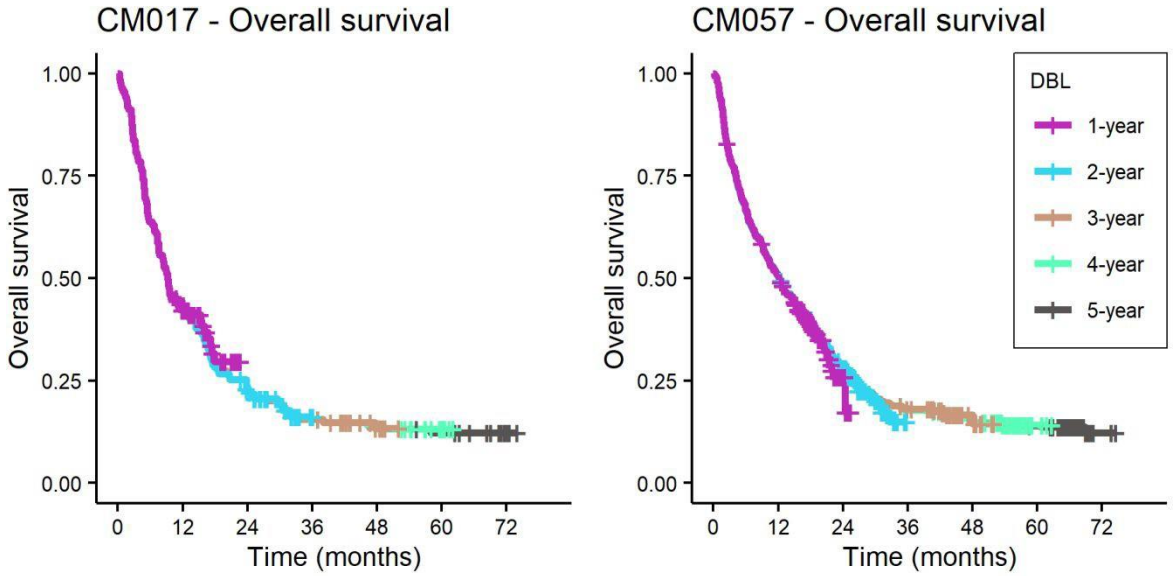
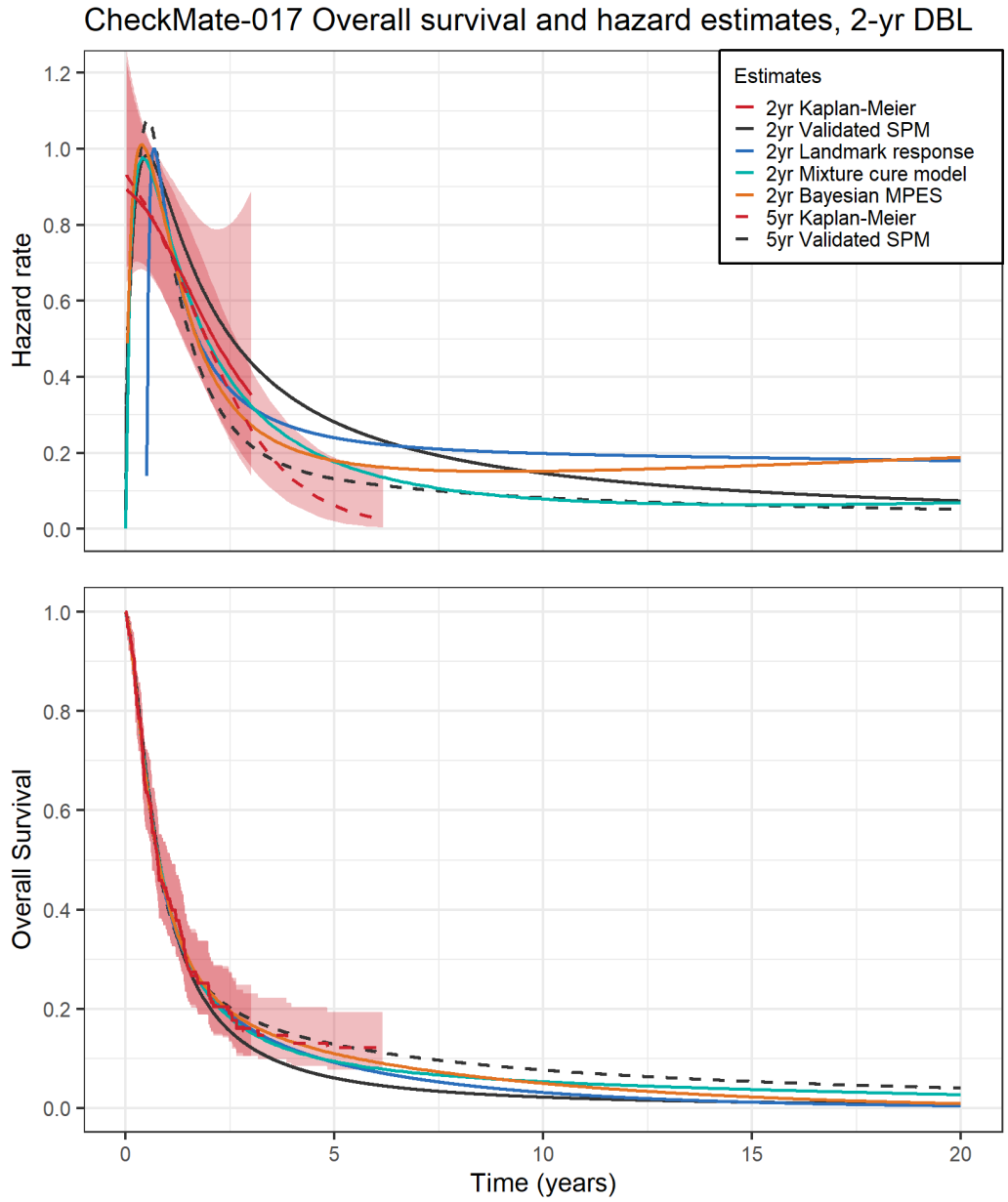
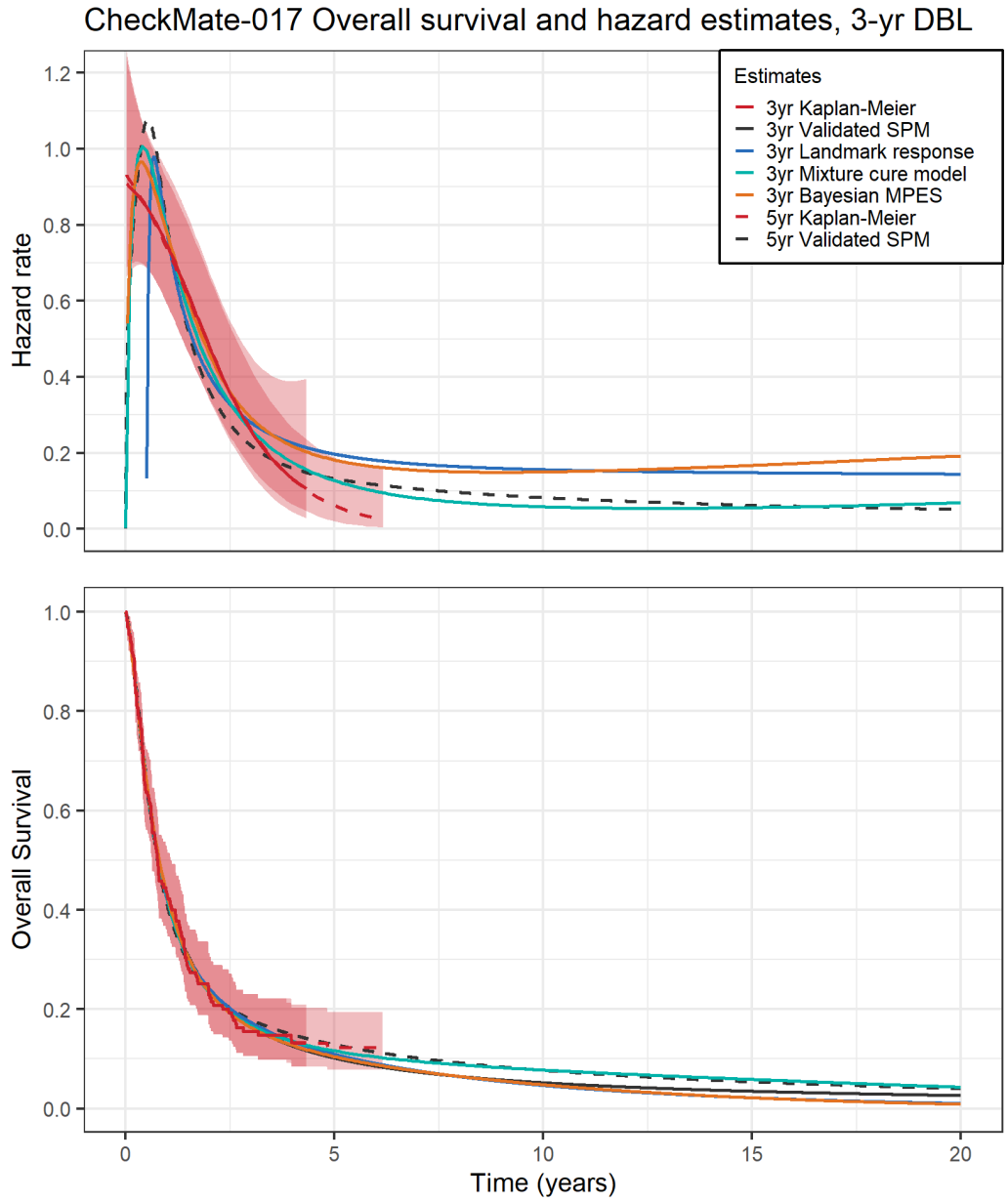


Figure 3: Overall survival probability, hazard rate, and restricted mean survival time (RMST) estimates for the nivolumab arm of CheckMate-017, 2-year database lock (DBL)



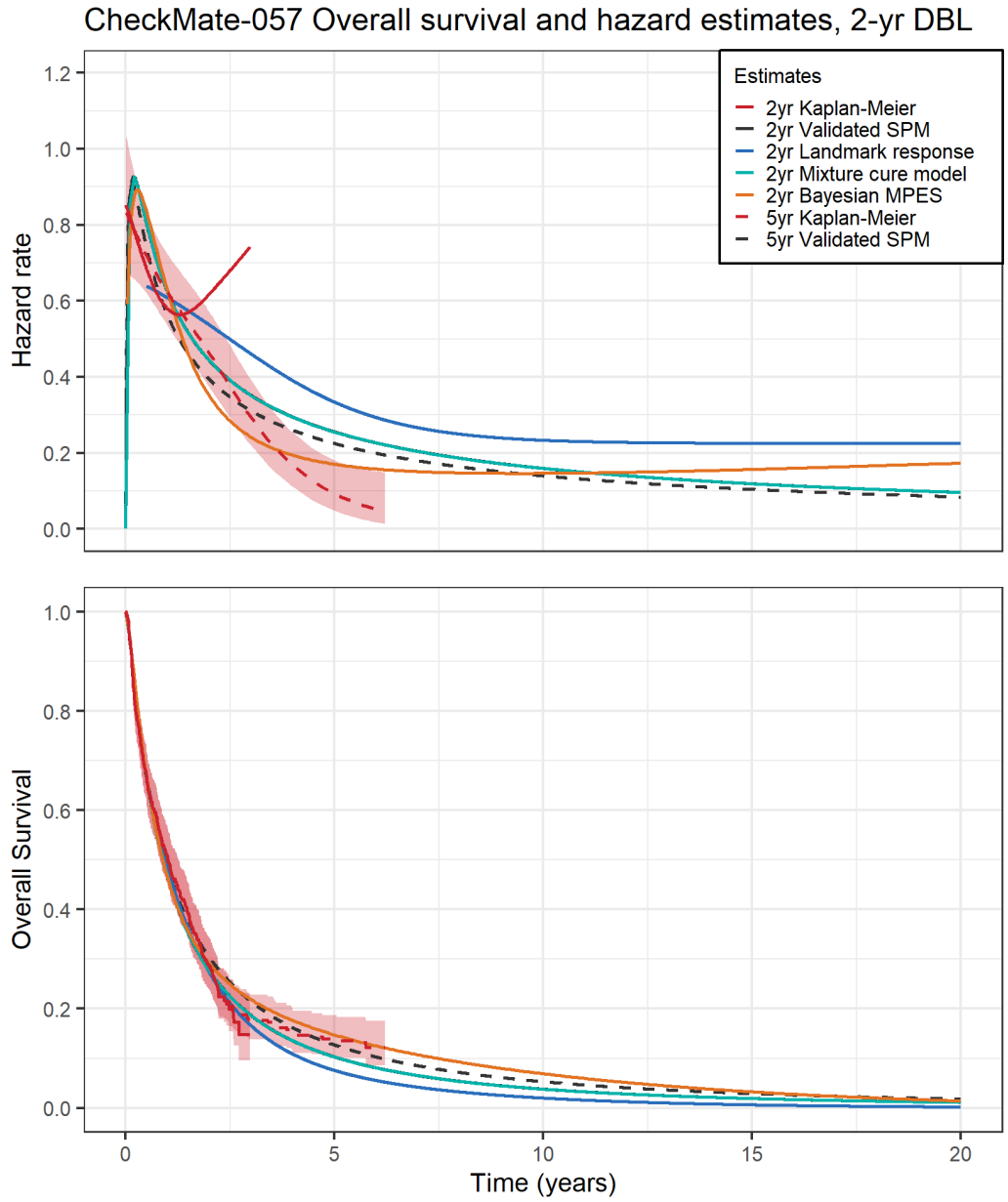
	Survival probability				RMST	
	5yrs	10yrs	15yrs	20yrs	5yrs	20yrs
2-yr Non-validated SPM	6.1%	2.2%	1.2%	0.8%	15.8	19.5
2-yr ERG SPM	3.8%	0.2%	0.0%	0.0%	16.0	16.7
2-yr Validated SPM	6.1%	2.2%	1.2%	0.8%	15.8	19.5
2-yr Landmark response	9.2%	3.2%	1.2%	0.5%	17.0	22.1
2-yr Mixture cure model	9.4%	5.3%	3.8%	2.7%	16.9	25.7
2-yr Bayesian MPES	11.0%	5.0%	2.3%	0.9%	17.5	25.0
5-yr Kaplan-Meier	12.3%	-	-	-	17.5	-
5-yr Validated SPM	12.9%	7.7%	5.4%	4.1%	17.9	30.5

Figure 4: Overall survival probability, hazard rate, and restricted mean survival time (RMST) estimates for the nivolumab arm of CheckMate-017, 3-year database lock (DBL)



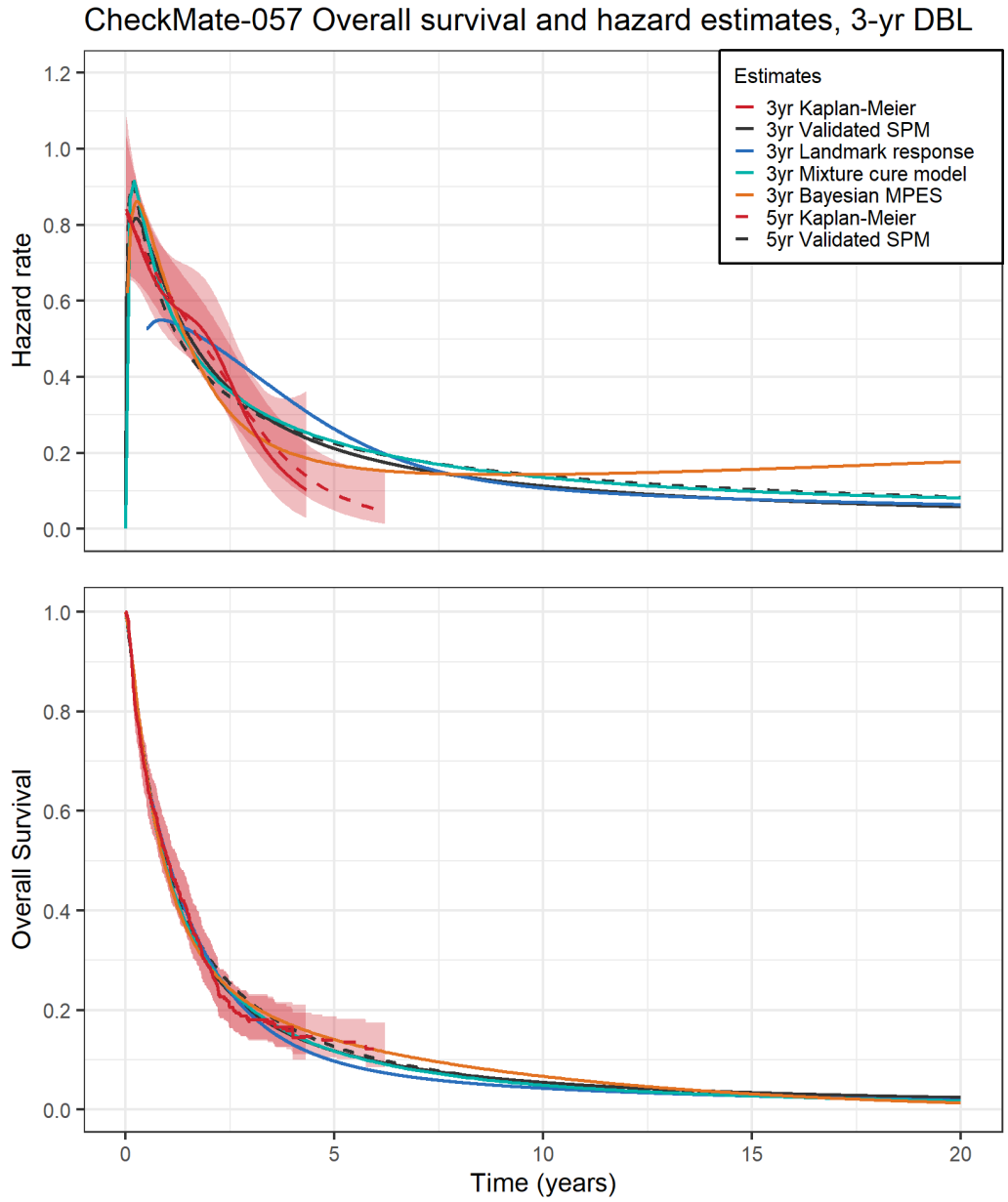
	Survival probability				RMST	
	5yrs	10yrs	15yrs	20yrs	5yrs	20yrs
3-yr Non-validated SPM	11.5%	5.7%	3.4%	2.2%	17.5	26.6
3-yr ERG SPM	6.2%	0.5%	0.0%	0.0%	17.5	19.0
3-yr Validated SPM	10.2%	5.2%	3.5%	2.6%	17.4	26.1
3-yr Landmark response	10.9%	4.7%	2.2%	1.1%	17.7	24.9
3-yr Mixture cure model	11.6%	7.8%	5.9%	4.3%	17.5	30.2
3-yr Bayesian MPES	10.5%	4.8%	2.2%	0.9%	17.4	24.6
5-yr Kaplan-Meier	12.3%	-	-	-	17.5	-
5-yr Validated SPM	12.9%	7.7%	5.4%	4.1%	17.9	30.5

Figure 5: Overall survival probability, hazard rate, and restricted mean survival time (RMST) estimates for the nivolumab arm of CheckMate-057, 2-year database lock (DBL)



	Survival probability				RMST	
	5yrs	10yrs	15yrs	20yrs	5yrs	20yrs
2-yr Non-validated SPM	3.5%	0.1%	0.0%	0.0%	17.3	17.9
2-yr ERG SPM	4.9%	0.3%	0.0%	0.0%	17.6	18.6
2-yr Validated SPM	10.3%	3.8%	1.9%	1.1%	18.7	25.0
2-yr Landmark response	7.6%	2.0%	0.6%	0.2%	18.0	21.4
2-yr Mixture cure model	10.3%	3.8%	1.9%	1.1%	18.7	25.0
2-yr Bayesian MPES	14.7%	6.9%	3.3%	1.4%	20.0	30.4
5-yr Kaplan-Meier	14.0%	-	-	-	19.4	-
5-yr Validated SPM	12.7%	5.3%	2.9%	1.8%	20.0	28.6

Figure 6: Overall survival probability, hazard rate, and restricted mean survival time (RMST) estimates for the nivolumab arm of CheckMate-057, 3-year database lock (DBL)



	Survival probability				RMST	
	5yrs	10yrs	15yrs	20yrs	5yrs	20yrs
3-yr Non-validated SPM	11.3%	4.9%	3.0%	2.1%	19.4	27.6
3-yr ERG SPM	7.5%	0.7%	0.1%	0.0%	19.3	21.2
3-yr Validated SPM	11.8%	5.5%	3.5%	2.5%	19.4	28.5
3-yr Landmark response	9.8%	4.3%	2.8%	1.9%	19.3	26.5
3-yr Mixture cure model	11.8%	4.9%	2.8%	1.8%	19.4	27.5
3-yr Bayesian MPES	14.1%	6.7%	3.2%	1.4%	19.8	29.9
5-yr Kaplan-Meier	14.0%	-	-	-	19.4	-
5-yr Validated SPM	12.7%	5.3%	2.9%	1.8%	20.0	28.6

References

1. Latimer, N. *NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data.* 2011 24th June 2021]; Available from: http://www.nicedsu.org.uk/

2. Rutherford, M., et al. *NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis.* 2020 24th June 2021]; Available from: http://www.nicedsu.org.uk/.

3. Gray, J., et al., *Extrapolation of Survival Curves Using Standard Parametric Models and Flexible Parametric Spline Models: Comparisons in Large Registry Cohorts with Advanced Cancer.* Medical Decision Making, 2020. **41**(2): p. 179-193.

4. Bullement, A., N.R. Latimer, and H. Bell Gorrod, *Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study.* Value Health, 2019. **22**(3): p. 276-283.

5. Othus, M., et al., *Accounting for Cured Patients in Cost-Effectiveness Analysis.* Value in Health, 2017. **20**(4): p. 705-709.

6. Ouwens, M.J.N.M., et al., *Estimating Lifetime Benefits Associated with Immuno-Oncology Therapies: Challenges and Approaches for Overall Survival Extrapolations.* PharmacoEconomics, 2019. **37**(9): p. 1129-1138.

7. Schadendorf, D., 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): p. 1889-94.

8. Borghaei, H., et al., *Five-Year Outcomes From the Randomized, Phase III Trials CheckMate 017 and 057: Nivolumab Versus Docetaxel in Previously Treated Non–Small-Cell Lung Cancer.* Journal of Clinical Oncology, 2021. **39**(7): p. 723-733.

9. Motzer, R.J., et al., *Nivolumab versus everolimus in patients with advanced renal cell carcinoma: Updated results with long-term follow-up of the randomized, open-label, phase 3 CheckMate 025 trial.* Cancer, 2020. **126**(18): p. 4156-4167.

10. Ascierto, P.A., et al., *Adjuvant nivolumab versus ipilimumab in resected stage IIIB-C and stage IV melanoma (CheckMate 238): 4-year results from a multicentre, double-blind, randomised, controlled, phase 3 trial.* Lancet Oncol, 2020. **21**(11): p. 1465-1477.

11. Topalian, S.L., et al., *Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer.* New England Journal of Medicine, 2012. **366**(26): p. 2443-2454.

12. Antonia, S.J., et al., *Four-year survival with nivolumab in patients with previously treated advanced non-small-cell lung cancer: a pooled analysis.* The Lancet. Oncology, 2019. **20**(10): p. 1395-1408.

13. Jackson, C., et al., *Extrapolating Survival from Randomized Trials Using External*

Data: A Review of Methods. Med Decis Making, 2017. **37**(4): p. 377-390.

14. Goeree, R., et al., *Economic evaluation of nivolumab for the treatment of second-line advanced squamous NSCLC in Canada: a comparison of modeling approaches to estimate and extrapolate survival outcomes.* J Med Econ, 2016. **19**(6): p. 630-44.

15. Rothwell, B., et al., *Cost Effectiveness of Nivolumab in Patients with Advanced, Previously Treated Squamous and Non-squamous Non-small-cell Lung Cancer in England.* PharmacoEconomics - Open, 2021. **5**(2): p. 251-260.

16. Chaudhary, M.A., et al., *Cost-effectiveness of nivolumab in squamous and non-squamous non-small cell lung cancer in Canada and Sweden: an update with 5-year data.* J Med Econ, 2021. **24**(1): p. 607-619.

17. Chaudhary, M.A., et al., *Cost-effectiveness of nivolumab in patients with NSCLC in the United States.* Am J Manag Care, 2021. **27**(8): p. e254-e260.

18. Smare, C., et al., *An economic evaluation of nivolumab for the treatment of squamous and non-squamous NSCLC in the Swedish setting.* Nordic Journal of Health Economics, 2019. **7**(1): p. 47-64.

19. Brahmer, J., et al., *Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer.* N Engl J Med, 2015. **373**(2): p. 123-35.

20. Borghaei, H., et al., *Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer.* N Engl J Med, 2015. **373**(17): p. 1627-39.

21. Amico, M. and I. Van Keilegom, *Cure Models in Survival Analysis.* Annual Review of Statistics and Its Application, 2018. **5**(1): p. 311-342.

22. Felizzi, F., et al., *Mixture Cure Models in Oncology: A Tutorial and Practical Guidance.* PharmacoEconomics - Open, 2021. **5**(2): p. 143-155.

23. Guyot, P., et al., *Extrapolation of Survival Curves from Cancer Trials Using External Information.* Med Decis Making, 2017. **37**(4): p. 353-366.

24. Royston, P. and Parmar, M. K., *Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects*. Statist. Med., 2002, **21**: 2175-2197.

25. Vickers, A., *An evaluation of survival curve extrapolation techniques using long-term observational cancer data*, Med Decis Making, 2019, **39**: p. 926-938.

26. Surveillance, Epidemiology, and End Results (SEER) Program, *SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2020 Sub (1975-2018) - Linked To County Attributes - Time Dependent (1990-2018) Income/Rurality, 1969-2019 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2021, based on the November 2020 submission.* www.seer.cancer.gov.

27. Carpenter, B., et al., *Stan: A probabilistic programming language*. J. Stat. Softw., 2017. **76**(1): p. 1-32.

28. Guo, J., et al., *RStan: the R interface to Stan. R package version 2.21.2*. 2020. https://cran.r-project.org/web/packages/rstan/

29. Hoffman, M. D. and Gelman, A.: *The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo*, J. Mach. Learn. Res, 2014, **15**: p. 1593-1623.

30. Royston, P. and Parmar, M. K. *Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome.* BMC Med. Res. Methodol., 2013. **13**: p 152.

31. Damuzzo, V., et al. *Analysis of survival curves: statistical methods accounting for the presence of long-term survivors*. Front. Oncol., 2019. **9**: p. 453.

32. Ash, B., Latimer, N. R., and Gorrod, H. B., *Survival extrapolation in cancer immunotherapy: a validation-based case study*. Value in Health, 2019. **22**(3): p. 276-283.