Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A Bayesian hierarchical mixture cure model demonstrated with data from the CheckMate 067 trial

Nathan Green*[1] | Murat Kurt[2] | Andriy Moshyk[2] | James Larkin[3] | Gianluca Baio[1]

[1]Department of Statistical Science, UCL, London, UK
[2]Worldwide Health Economics and Outcomes Research, Bristol Myers Squibb, NJ, USA
[3]The Royal Marsden Hospital, London, UK

Correspondence
*Nathan Green Email: n.green@ucl.ac.uk

Summary

Time to an event of interest over a lifetime is a central measure of the clinical benefit of an intervention used in a health technology assessment (HTA). Within the same trial multiple end-points may also be considered. For example, overall and progression-free survival time for different drugs in oncology studies. A common challenge is when an intervention is only effective for some proportion of the population who are not clinically identifiable. Therefore, latent group membership as well as separate survival models for groups identified need to be estimated. However, follow-up in trials may be relatively short leading to substantial censoring. We present a general Bayesian hierarchical framework that can handle this complexity by exploiting the similarity of cure fractions between end-points; accounting for the correlation between them and improving the extrapolation beyond the observed data. Assuming exchangeability between cure fractions facilitates the borrowing of information between end-points. We show the benefits of using our approach with a motivating example, the CheckMate 067 phase 3 trial consisting of patients with metastatic melanoma treated with first line therapy.

KEYWORDS:
Bayesian, survival analysis, oncology

1 | INTRODUCTION

Interventions that impact the time to an event of interest, such as disease progression or death in a cancer trial, form a high proportion of appraisals by Health Technology Assessment (HTA) agencies, such as the National Institute for Health and Care Excellence (NICE), in England[1]. For instance, a new intervention may be expected to increase the amount of time until a patient experiences disease progression or death from any cause. For this reason, "survival data" are instrumental to modelling in HTA[2,3].

A feature of survival data from a trial is that they are often reported on during as well as at the end of the trial period. This enables on-going assessments and can inform decisions such as an early end to the trial due to harm or futility. However, in reality, it is normally not possible to make use of the collected data in real time because there is a period in which they need to be cleaned due to issues, such as missingness and other data quality errors. Therefore, snapshots of the trial data are made at pre-specified points defined in the study protocol. This can be at a particular date, follow-up time, or a particular number of patients. These subsets of survival data made at discrete times are called *data-cuts*.

[0]**Abbreviations:** MCM, mixture cure model

In order to quantify accurately the health and economic benefits of a new intervention using such survival data, it is necessary to estimate the *mean survival time*, ie, the *long-term* effects of a given intervention. This is in contrast to general survival curve summaries, such as the median time, which are typically used in standard "biostatistical" analyses. This idiosyncrasy has important implications because, in order to calculate the mean survival time we require the full survival curve, ie, over a patient's lifetime, but available data (e.g. data-cuts from a randomised trial) almost inevitably only cover a limited time frame and are subject to a high degree of censoring. Thus, it is necessary to *extrapolate* the observed survival curve to a time horizon that is typically much longer than the one considered in the experimental study. Consequently, unlike "standard" time-to-event analyses that are often based on semi-parametric models (e.g. Cox regression), HTA modelling is based on a fully parametric approach to the survival analysis, as recommended by a highly influential NICE Decision Support Unit (DSU) Technical Support Document (TSD)[1].

Another important feature of survival data in the field of cancer research has arisen in the past decade due to the development of potentially highly successful immuno-oncology treatments. These aim to stimulate the body's immune system to recognise and kill cancer cells[4]. Several types of immunotherapy can be used to treat melanoma including immune checkpoint inhibitors, interleukin-2 (IL-2), and oncolytic virus therapy. In particular, inhibitory immune checkpoint blockade combination therapies have shown significant success in promoting immune responses against cancer and can result in tumour regression in many patients[5]. These advances have lead to cancer patients with improved survival end-points and more long-term survival (LTS) (although complete responders, CR, still have worse outcomes than the general population). In previous immuno-oncologic studies for melanoma therapies, such as those evaluating ipilimumab and nivolumab, results have indicated that survival curves "plateau" for a considerable proportion of LTS patients in the time period covered by survival curve[6,7]. That is, the underlying survival curve converges to a probability (substantially) greater than zero and does not appear to decrease further. Note that in practice, in the tail of the Kaplan-Meier plateauing may also be due to sampling issues, such as small sample size or missing data, and so there will be uncertainty about the true proportion of LTS. When a survival curve exhibits this plateau behaviour, it is common to adopt a *mixture cure model* (MCM) approach to describe the generation of the underlying data. A MCM considers a population as a mixture of two groups: the cured and not cured. In our case, the cured group contains the LTS patients. In many MCM analyses the survival associated with the uncured fraction of the population is represented by a Cox proportional hazard model.

However, the survival plateaus indicating a proportion of LTS may not have been reached at the time of a data-cut (including the planned end of the study). In order to obtain complete survival curves to use in an HTA evaluation we need to extrapolate the survival curves to the sustained plateaus. Ideally, these extrapolations should be consistent between data-cuts[8]. The question of how to principally and usefully do this in situations where trials include multiple treatment arms and end-points is the subject of this paper. The fact that we consider multiple end-points within a trial in the same analysis means that information about the plateau from data for one end-point can potentially help inform the estimation of the plateau for another end-point. This can enable consistency between plateau estimates and maximise the utility of all available data.

Within a frequentist paradigm, an example of related work is a multi-level modelling approach in mixture cure modelling using random effects to model multilevel clustering structure in the linear predictors in both hazard function and cured probability parts[9]. Correlation between random effects in the uncured survival and cure fraction has been investigated using a bivariate Normal distribution[10 11]. Instead, we will impose dependency with hyperparameters. To our knowledge there has been no Bayesian fully-parametric mixture cure model with a multi-level modelling structure for end-points.

The paper is structured as follows. In the next section, the motivating example and data used throughout is outlined. We introduce the basic mixture cure model in Section 3.1, before extending this to our Bayesian hierarchical model in Section 3.2. In Section 4 our novel modelling approach is applied to the motivating example dataset. Comparisons are made between our approach and an independent model analogue. Finally, in Section 5 we discuss the results and future work.

2 | MOTIVATING EXAMPLE

Our motivating example concerns a long-term study of melanoma therapies. The study has been described in detail elsewhere, so we give a brief overview of the data here (see[6,7,12] for more details). The *CheckMate 067* trial is a phase 3 randomised, double-blind study conducted on eligible patients aged 18 years or older with previously untreated, unresectable stage III or IV melanoma. Nivolumab alone or the combination of nivolumab plus ipilimumab was compared with ipilimumab alone in patients with metastatic melanoma. They were randomly assigned in a 1:1:1 ratio and stratified by anti–programmed death-1 ligand 1

(PD-L1) expression status, whether there is a mutation in the BRAF gene that makes it work incorrectly (BRAF mutation), and AJCC metastasis stage. The total baseline sample consists of $n = 945$ subjects divided across three treatment arms i) ipilimumab monotherapy with 3 mg/kg IV once every 3 weeks (Q3W) for a total of 4 doses; ii) nivolumab monotherapy with 3 mg/kg intravenous (IV) once every 2 weeks (Q2W); iii) combined nivolumab with ipilimumab with 1 mg/kg IV and 3 mg/kg IV Q3W, respectively, for 4 doses followed by nivolumab 3 mg/kg IV Q2W. The co-primary end-points were *progression-free survival* (PFS; ie, the time from randomisation until the first of disease progression or death) and *overall survival* (OS; ie, the time from randomisation to time of death). The data source for the analysis consists of patient-level data of times to PFS and OS with covariates sex, age at start of trial, and country of centre.

In our data set the minimum study follow-up from randomisation of 60 months was used as the final data-cut time. We also considered two earlier artificial data-cuts by truncating the 60-month data at 12 and 30 months. These particular times were chosen for this analysis because the actual study had a planned data-cut at 28 months, and the largest observed median time for OS was just over 30 months. Thus, this time is reasonable and relevant to the trial, and at this time it may be possible to make comparable inferences to the complete 60-months data set but at a significantly earlier time. Moreover, the latest median time between all treatment regimens for PFS was approximately 12 months. Therefore, at 12 months it may be possible to characterise the PFS survival curves but not the less mature OS and so borrowing information between PFS and OS may be advantageous. Taking a data-cut at an even earlier time would probably not provide sufficient information to enable any useful extrapolation. Patients who were still being followed-up but had not experienced the event of interest at the time of an artificial data-cut were censored at this time.

The reader is encouraged to refer to previously published analyses for the CheckMate 067 trial results, including the 5-year database lock (DBL) publication[7]. In summary, for PFS the median (95% CI) times to event for each treatment arm were 2.86 (2.79-3.29) for ipilimumab, 6.93 (5.32-10.41) for nivolumab, and 11.50 (9.26-20.80) for the combination treatment of nivolumab and ipilimumab. For OS the median (95% CI) times to event for each treatment arm were 20.0 (17.22-25.6) for ipilimumab, 36.9 (31.24-60.9) for nivolumab, and not observed for the combination treatment. Figure 1 shows the Kaplan-Meier estimates of the PFS and OS survival curves for the complete CheckMate 067 phase 3 trial data. Both sets of curves, for PFS and OS, appear to approach a positive limiting probability; the PFS curves more clearly so. Early published results for outcomes at 7.5 years follow-up continue to demonstrate the durability of responses with combined nivolumab and ipilimumab, and an ongoing survival plateau[13].

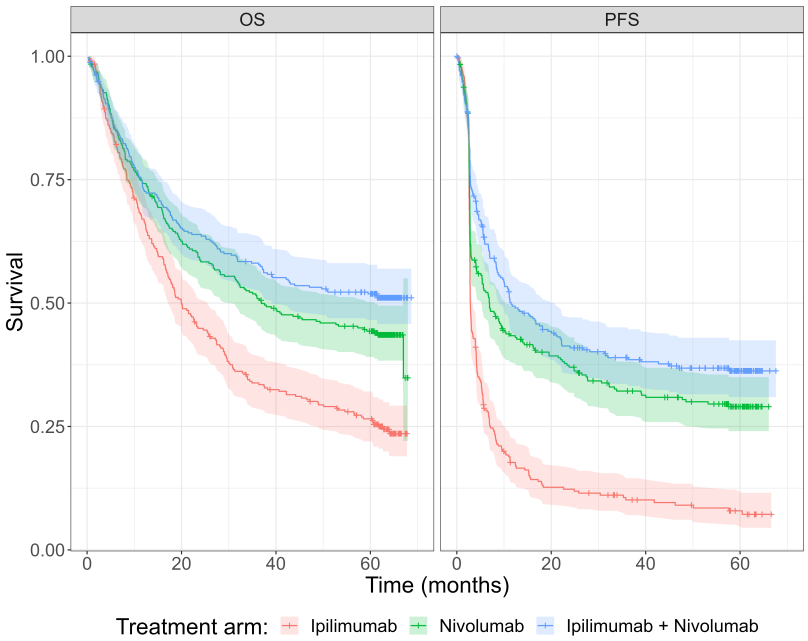

FIGURE 1 Kaplan-Meier curves of PFS and OS times with 95% CI for the complete CheckMate 067 phase 3 trial data and ipilimumab, nivolumab and combination treatments.

A standard frequentist analysis has been carried-out on these data already with separate mixture cure models for each treatment and OS or PFS [14]. It was estimated that for OS, ranges of estimated proportions of LTSs were 16%–26% for ipilimumab, 38%–46% for nivolumab, and 49%–54% for combined treatment across all modelled distributions. Similarly, for PFS, ranges of the proportions of LTSs were 9%–13% for ipilimumab, 29%–33% for nivolumab, and 38%–40% for combined treatment.

3 | MODELLING FRAMEWORK

Consider a study in which n individuals are observed, and the (non-negative) time to either progression (T_i^{prog}) or death (T_i^{death}) are recorded for each individual $i = 1, \dots, n$. Denote the event time for individual i by T_i. In our motivating example, we would have a separate set of T_i for either end-point, PFS or OS. Thus, PFS times are $T_i = \min(T_i^{prog}, T_i^{death})$, and OS times are $T_i = T_i^{death}$. Event times may be censored, and therefore we consider a censoring time C_i and an event indicator $\delta_i = \mathbb{I}(T_i < C_i)$ — so that $\delta_i = 1$ if the end-point is fully observed for individual i and 0 if it is censored. The observed survival time is then indicated as $t_i = \min(T_i, C_i)$.

Typically, a study will also measure some individual-level covariates, which we assume are collected in a vector $\boldsymbol{x}_i = (x_{1i}, \dots, x_{Pi})^\top$. These may include the individual's age, sex, comorbidities, and usually, a treatment arm indicator. The full data can then be represented as $\mathcal{D}_i = (t_i, \delta_i, \boldsymbol{x}_i)$. In our example, the covariates are the same between PFS and OS.

We model the observed times using a parametric sampling distribution $t_i \sim p(t_i \mid \boldsymbol{\theta})$ as a function of a set of model parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}(\boldsymbol{x}), \mu(\boldsymbol{x}))$. In this notation, the location parameter $\mu(x)$ indicates the mean or the scale of the probability distribution; and a (set of) ancillary parameters $\boldsymbol{\alpha}(x)$ describe the shape or variance of the distribution. Whilst it is possible for both μ and $\boldsymbol{\alpha}$ to explicitly depend on the covariates $\boldsymbol{x}$, usually the formulation is simplified to assume that these only affect directly the location parameter. Since $t > 0$, we typically use a generalised linear formulation. So, for an individual i

$$g(\mu) = \beta_0^\mu + \sum_{p=1}^{P} \beta_p^\mu x_{pi} \; [+ \dots],$$

to model the rate or location parameter, where β_0^μ represents an intercept, and β_p^μ are the coefficients for an individual's measurements. The $[+\dots]$ term indicates additional components for the linear predictor (e.g. structured random effects). The function $g(\cdot)$ is typically the logarithm.

3.1 | The standard mixture cure model

Suppose that we observe a sustained plateau in the survival time data or have some other reason to believe that there is a long-term survivor group in our sample. We will now present the standard mixture cure model (sometimes called a long-term survival model) as the basis of our proposed extension. Denote the probability of being cured by the *cure fraction*, π. The term "cure" here refers to the type of statistical model and does not literally mean that they are cured of disease. In our case, cured means that a patient has responded to treatment and is a LTS. Because the cure models explicitly split patients in to two groups (cured or not) then the full set of model parameters consists of two separate groups. Define $\boldsymbol{\omega} = (\boldsymbol{\theta}^b, \boldsymbol{\theta}^u)$, where superscript b denotes the *background* model and u denotes the *uncured* model. In the standard cure model, cured individuals are usually assumed to have no chance of experiencing the event of interest in the period under consideration, simplifying the model. There are several types of cure models and a rich and growing body of research in this area. (see [15] for a comparison and guidance with application to oncology.)

For an individual i in a population with a cure fraction, the mixture cure model is associated with the following survival function

$$S(t_i \mid \boldsymbol{\omega}, \boldsymbol{x}_i) = S_b(t_i \mid \boldsymbol{\theta}^b, \boldsymbol{x}_i) \left[\pi + (1-\pi) S_u(t_i \mid \boldsymbol{\theta}^u, \boldsymbol{x}_i)\right], \tag{1}$$

where $S(t) = 1 - \int_0^t p(s \mid \theta)\, \mathrm{d}s$ denotes survival at time t, $S_b(t \mid \boldsymbol{\theta}^b, \boldsymbol{x})$ is a function of the background mortality at time t conditional on covariates $\boldsymbol{x}$, and $S_u(t \mid \boldsymbol{\theta}^\mu, \boldsymbol{x})$ is a function of the (excess) mortality due to cancer at time t, conditional on covariates $\boldsymbol{x}$. The formulation in equation (1) can be thought of as comprising of two submodels: i) an *incidence* model for the cure fraction π (the general term from epidemiology is the rate or probability of a disease); and ii) a *latency* model for the survival function of the uncured population S_u. This is 'hidden' since it cannot be directly observed from the data.

The model above can be extended to account for multiple treatments, K; for instance, in our running example, $K = 3$ and we could specify the cure fraction for treatment $k = 1, \ldots, K$ using the following "fixed" effect model

$$\text{logit}(\pi_k) = \beta_k^{\pi} \, [+ \ldots], \tag{2}$$

where β_k^{π} are the regression coefficients quantifying the impact of treatment k. The $[+ \ldots]$ term could include a frailty term in the model. The treatment index for an individual is included in the covariate vector $\boldsymbol{x}$ so that π is replaced by $\pi(\boldsymbol{x})$ in equation (1). Note that although we are including all treatments arms in the same model we are not assuming any relationship between them. In particular, we do not suppose proportional hazards. This allows greater flexibility for the cases when proportional hazards is violated.

3.1.1 | Issues modelling multiple end-points

There are several potential issues with existing approaches to cure modelling, when applied to an HTA context in particular.

Firstly, when the end-points are correlated, such as with the common events of interest overall survival (OS) and progression-free survival (PFS), this ought to be accounted for in the modelling. There may be information in one sample of event times that we can use to improve the inference from another. Performing separate, independent analyses may even give counter-intuitive results. In the case of multiple types of end-points in the same trial, since the clinical interpretation of a cure fraction is the proportion of patients that are LTS, it may make intuitive sense that there is a single underlying cure fraction. For example, what would be the interpretation if there are differences in emergent survival plateaus between PFS and OS? This is a clinically unintuitive dichotomy between the resulting proportions of long-term survivors.

Secondly, time to event data are typically right-censored due to administrative censoring or loss to follow up. The censoring times may be too early in order to adequately characterise a survival model for HTA for a single end-point. In particular, OS times are bounded below by PFS times by definition, so will be subject to more censoring at the same cut-point. By taking advantage of the set of end-points together, the data paucity may be alleviated to give better fits and extrapolation to the full lifetimes of patients. This is particularly important for health economics modelling in order to gauge the full cost and effect of an intervention.

Furthermore, if there is heavy censoring for all endpoints, such as early data cuts, and the number of endpoints is small, then the data alone may not represent the true underlying survival curves, including the plateau for the cure fractions. At this point, additional sources of information should be principally incorporated into the model. A Bayesian paradigm naturally allows for the inclusion of external evidence from sources such as historical data and expert opinion. It is reasonable to assume, in the absence of additional information, that the cure fractions for all end-points are drawn from the same or very similar distribution. In order to do this we can increase the amount of strength to borrow between end-point data which strikes a balance between the information (or lack of) in the data and other knowledge[16]. Practically, this corresponds to a reduction in the global cure fraction variance hyper parameter which in the limit is a base model with $\sigma_k = 0$, equivalent to complete pooling of the data.

We will now present a Bayesian hierarchical mixture cure model for multiple end-points. This has the double advantage of borrowing information between end-points (e.g. in the case of OS and PFS, the likely more mature PFS data can inform the often highly censored OS analysis) and obtaining a single cure fraction estimate.

3.2 | The hierarchical mixture cure model

In this section, we show how to construct the extended mixture cure model in order to model multiple end-points together and alleviate the issues posed in the Section 3.1.1. In order to do this, we must first introduce an additional index to denote an end-point $j = 1, \ldots, J$ — typically, $J = 2$ for PFS and OS. We can now define a separate mixture cure model, as described in equation (1), for each of the end-points j. With this in mind, let us consider three alternative multiple end-point incidence models for the estimation of the cure fraction for treatment k and end-point j, denoted π_{kj}.

Firstly, the approach taken so far, and presented in Section 3.1, is to model the cure fraction corresponding to each end-point completely separately (no pooling), assuming that they are independent.

$$\pi_{kj} \perp\!\!\!\perp \pi_{kj'}, \; j, j' = 1, \ldots, J.$$

Secondly, we can assume that the cure fraction is the same for all end-points and so pool (or "lump" together) the samples (complete pooling). In our example, PFS can be used as a proxy for OS since there are fewer missing data. There may be no reason why the cure fraction should be different between different end-points, especially if individuals are observed for long

enough. In this case

$$\pi_{kj} = \pi_{kj'},\ j, j' = 1, \ldots, J.$$

A potential issue with the complete pooling approach is that it ignores any variation in cure fraction that exists between end-points.

Lastly, and the approach proposed in this paper, is a compromise between the first two approaches called partial-pooling. We propose a hierarchical structure on the cure fraction assuming exchangeability between all end-points j for each treatment k. We assume that there is a single, shared "common" cure fraction from which the separate end-point specific cure fractions are obtained, corresponding to equation (2) but where π_k, β_k^{π} are no longer for particular end-points as before but now represent "global" values. Using the global cure fraction we can define the second-level end-point specific cure fractions as

$$\text{logit}(\pi_{kj}) \sim \text{Normal}(\nu_k, \sigma_k^2),\ j = 1, \ldots, J. \tag{3}$$

where $\nu_k = \text{logit}(\pi_k)$ and σ_k^2 is the between-group (end-points) variance. The end-point specific cure fractions in our example are denoted π_{PFS}, π_{OS}. Figure 2 shows a directed acyclic graphical (DAG) representation of the Bayesian hierarchical mixture cure modelling framework for two end-points. The equivalent DAGs for the complete and no pooling models are given in the Appendix. Note that pooling refers to the cure fractions. The survival times are not pooled; the parameters for the uncured survival models (ϕ, β^{λ} in the latency model) are estimated using the separate sets of survival times for each end-point eg, in the complete pooling case we do not assume that PFS and OS have the same survival time distributions. However, in the complete and partial pooling case there is an indirect pooling effect for the latency models via the incidence model, so the fact that the cure fractions are similar/the same will influence the shapes of the survival curves.

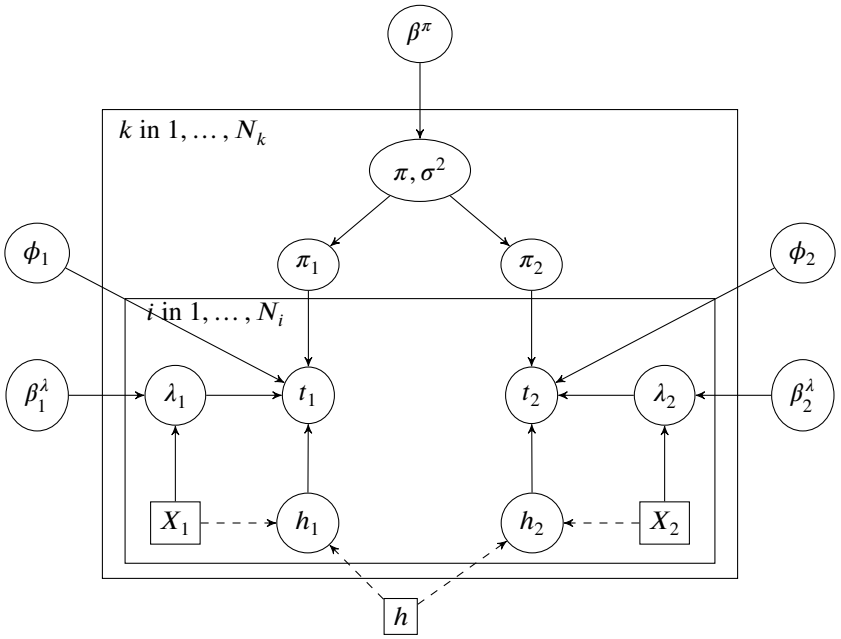

FIGURE 2 A hierarchical mixture cure model DAG for a trial with two end-points. The trial indices correspond to the motivating example end-points of PFS (1) and OS (2). Solid lines represent stochastic and dashed lines deterministic relationships, respectively. Cured patients have fixed hazards e.g. taken from life tables. The distribution of times for uncured patient regresses on covariates for the rate parameter and ϕ is the set of ancillary parameters.

4 | APPLICATION

In this section we apply the models from Section 3.2 to the CheckMate 067 dataset. We first detail the modelling aspects specific to this analysis before presenting the results.

This analysis was carried-out using the Stan inference engine [17] called from R v.4.3.1 [18] using the package `rstan` on a Windows 10 PC. Each fit employed 4 chains of 1000 iterations each, and a burn-in of 100 iterations. In order to determine convergence, we checked effective sample sizes, Markov chain standard error and performed posterior predictive checks. Details of the algorithm and model checking can be found in the Appendix. The code has been developed into an R package using a generalisable framework. This means that analyses for an arbitrary number of end-points can be easily performed with new data. The package is freely-available and can be downloaded from https://github.com/StatisticsHealthEconomics/multimcm.

4.1 | Background survival

In many cure fraction models, it is common to have $S_b(t, x) = 1$ for all t under consideration, which is reasonable in the short-term. However, we are focused on the full life-course of an individual in an HTA context and so include a general population all-cause mortality competing risk. Thus, the uncured population is subject to a background mortality risk in addition to the risk from the event of interest.

We used the World Health Organisation (WHO) life tables by country for the latest year available of 2016 [19] to inform the background mortality in the mixture cure model. The baseline hazards are the expected mortality rate for each patient at the age at which they experience the event. The mortality data are adjusted for country, age and gender. Thus, this structure provides a granular account of the different patient profiles in the trial. The WHO reports conditional probabilities of death in 5-year intervals until age 85. A constant annual mortality rate is reported for individuals over 85. They assumed that no-one lives beyond 100 years. Further details are provided in the supplementary material.

In a Bayesian analysis, there are alternative ways in which we could model the background mortality. For this work we used WHO hazard point estimates as fixed and known. We could consider the WHO estimates to provide sufficiently accurate estimates given the sample size, and so incorporating uncertainty is less crucial than in other cases. This also forces consistency across fits. Denote the fixed WHO estimates for individual i used for the background survival in equation (1) as $S_b(t_i \mid \boldsymbol{\theta}_b, \boldsymbol{x}_i) = \hat{S}_i$.

It may be that the patients in the trial have a worse background survival than the average equivalent individual in the general population, and so naively using the WHO life tables will over-estimate their survival times. There are several possible ways in which this could be addressed. For example, using only the complete responders in the sample, who are clinically confirmed as cured, it is possible to obtain a posterior distribution for a hazard ratio between these and a WHO estimate baseline. This can serve as a prior in the main model in a two-step approach. Alternatively, prior belief could be defined explicitly using expert knowledge. This could be elicited directly for the hazard ratio or on a natural scale, such as mean lifetime, and transformed.

4.2 | Prior specification

For the latency model, we centred covariates to aid computation and specified vague priors on the log-scale for the coefficients of the OS and PFS rates, $\log(\lambda_{OS})$ and $\log(\lambda_{PFS})$, respectively. In the baseline scenario these were: $\beta^{\lambda}_{PFS,0} \sim \text{Normal}(-3, 0.5)$, $\beta^{\lambda}_{OS,0} \sim \text{Normal}(-3, 0.5)$ and for age $\beta^{\lambda}_{PFS,age} \sim \text{Normal}(0, 0.01)$, $\beta^{\lambda}_{OS,age} \sim \text{Normal}(0, 0.01)$. The prior distributions for the ancillary parameters ϕ_{OS} and ϕ_{PFS} depended on the particular survival distribution used. For log logistic and Weibull the prior distribution for the shape parameter was $\phi \sim \text{Gamma}(1, 1)$, for Gompertz $\phi \sim \text{Gamma}(1, 1000)$, and for log Normal the prior distribution for the standard deviation was $\phi \sim \text{Gamma}(1, 2)$.

For the incidence model, the global cure fraction was modelled as a fixed effect linear regression with logistic link. Each treatment coefficient prior is independent $\beta^{\pi}_{k} \sim \text{Normal}(-0.1, 0.2)$. This corresponds to a prior mean cure fraction of just under 0.5, and approximately 10% and 1% chance of exceeding 0.6 and 0.7 respectively. Alternatively, we could have specified the cure fraction on the natural scale between 0 and 1 using a Beta prior distribution $\pi \sim \text{Beta}(a, b)$. This has the advantage of representing the probability directly which could aid elicitation but has the disadvantage of a and b being less interpretable. The Beta distribution parameters can always be obtained via transformation of the mean and standard deviation. Prior predictive plots and further discussion are provided in the supplementary material.

4.2.1 | Cure fraction between-endpoint variance

For the complete data analysis, the random effect variance on the global cure fraction was modelled using a vague (folded) half-Normal [20] with $\sigma_k \sim \text{Normal}(0, 2.5^2)\mathbb{I}(0,)$, where $\mathbb{I}(0,)$ denotes a distribution truncated at mean 0. This was found to be the best option for this study. As a rule of thumb, when the number of groups is greater than 5 then a minimally informative uniform

prior density associated with the standard deviation σ is expected to generally work well, or when there is small hierarchical variance then a Gamma(2, 0.1) may be preferred[21].

As discussed, the early phase of a trial is not very informative of the cure fraction if plateauing is yet to occur. In this situation, we can take advantage of the Bayesian hierarchical structure. In the absence of other evidence, it is reasonable to assume that uncertainty about the cure fractions for each end-point can be modelled using a common distribution, which corresponds to a narrow prior on the between-group standard deviation at the global cure fraction level. In order to do this, we modelled the global cure fraction with a *Penalised Complexity (PC)* prior[22] that penalizes departure from a base model and preferences simplicity.

In practice, the PC prior approach places an exponential distribution on σ_k in equation (3). The rate parameter for this distribution, ρ, is obtained by first specifying a threshold value σ_0 and probability of exceeding that thereshold, α, ie, $p(\sigma > \sigma_0) = \alpha$. These can then be plugged into $\rho = -\log(\alpha)/\sigma_0$, to obtain the rate ρ of the exponential distribution prior. For specific values, for $p(\pi_k > 0.55) = 0.005$ then $\sigma_0 = 0.18$ and $\rho = -\log(0.01)/0.18 = 25$ or $\sigma \sim \text{Exp}(25)$. Similarly, for $p(\pi_k > 0.525) = 0.005$ then $\sigma_0 = 0.08$ and $\rho = -\log(0.01)/0.08 = 55$ or $\sigma \sim \text{Exp}(55)$. The latter prior was used for the data-cut at 12 months. We can see how these two prior distributions correspond to the prior uncertainty on the cure fraction scale in Figure 3. The two distributions are similar, but the plot for $\rho = 55$ is more peaked around the central value, preferencing the base model even more.

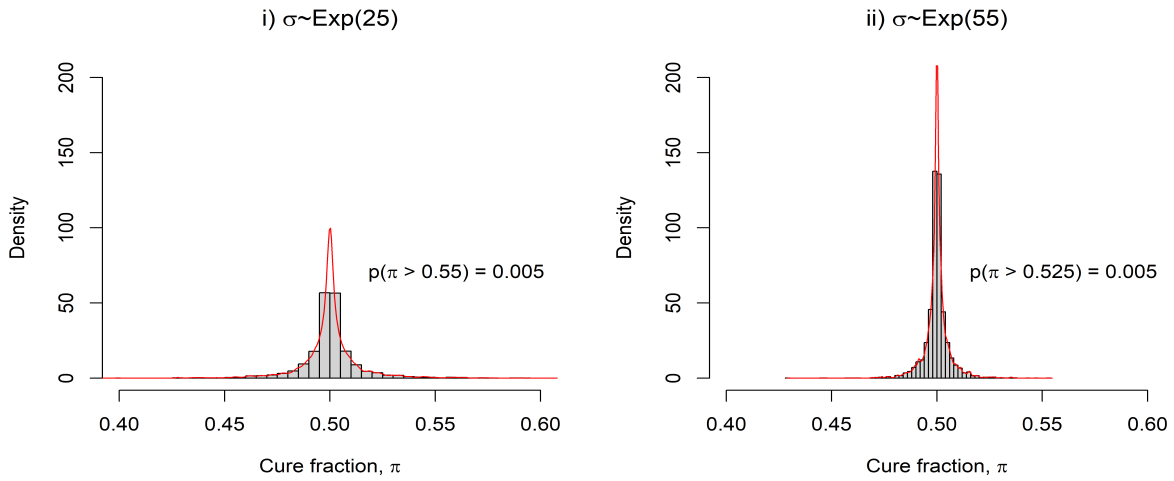

FIGURE 3 Penalised complexity prior distributions on the cure fraction scale corresponding to different between-group variances at the global cure fraction level of the Bayesian hierarchical mixture cure model.

4.3 | Parametric models for the sampling variability

To model the sampling variability in the observed OS and PFS event times we used the exponential, Weibull, Gompertz, log-normal and log-logistic distributions. These are among those recommended by NICE guidelines for Health Technology Assessments (HTA)[1]. In general, providing this list has been interpreted by many in the HTA literature as meaning all of these distributions should be modelled with a given data set. This prescriptive approach does not take into account what may be known *a priori* about the problem. In reality, a subset of these parametric models will be more plausible for the data and so these are the ones that should be investigated.

4.4 | Model assessment

We evaluated the goodness of fit of the overall models using out-of-sample predictions estimated with the widely applicable information criterion (WAIC) (and leave-one-out cross-validation, LOO, given in the Appendix)[23]. These have various advantages to more common AIC and DIC. WAIC is the difference between the computed log pointwise predictive density and the estimated effective number of parameters. Smaller values are better. It is fully Bayesian and invariant to parameterisation. They

are easily obtained using the posterior sample of the Stan output. The set of parameters in the hierarchical model (and their dimensions) are $\boldsymbol{\beta}^{\boldsymbol{\pi}}$ (3), $\boldsymbol{\sigma}$ (3), $\boldsymbol{\beta}^{\boldsymbol{\lambda}}_{\mathbf{PFS}}$ (2), $\boldsymbol{\beta}^{\boldsymbol{\lambda}}_{\mathbf{OS}}$ (2), $\boldsymbol{\phi}_{\mathbf{OS}}$ (1), $\boldsymbol{\phi}_{\mathbf{PFS}}$ (1), $\boldsymbol{\pi}_{\mathbf{OS}}$ (3), $\boldsymbol{\pi}_{\mathbf{PFS}}$ (3).

There has not been much work carried-out to date on model assessment methods specific to mixture cure models. For the basic mixture cure model, assessment may focus on the model as a whole or on the incidence and latency submodels separately. The fit of the latency submodel can be examined using a revised Schoenfeld residual[24], and the performance of the incidence submodel can be checked with concordance measures such as AUC[25]. However, these incidence model approaches are best used directly with known cure status and are designed for models which consider individual cure fractions, unlike our model which assumed a single population cure fraction. Of course, we also extend the basic mixture cure model so it is not yet clear how to best use these methods in this situation.

4.5 | 60 months data only results

In this section, we will explore the mixture cure model fits using the complete 60-months CheckMate 067 trial data. An example of the Bayesian hierarchical model posterior survival curves using the complete data is shown in Figure 4. Equivalent plots for all other distributions and a combined plot are given in the Appendix.

The background survival is linearly decreasing with time and the same between plots, as expected because were are using point estimates from WHO life tables. Visual inspection demonstrate reasonably good model fits. The PFS Kaplan-Meier curves show an initial steep decrease before levelling off. The exponential fits particularly well to the tail which is of particular importance.

All model fits are fairly similar. For example, for the exponential model, the PFS median times are 5, 9 and 12 months, for ipilimumab, nivolumab and combined treatments, respectively. In comparison with the Kaplan-Meier median times, the fitted times are 75%, 30% and 5% larger. That is, the median times converge for large times. Similarly, for the exponential model, the OS median times are 22, 38 and 56 months, for ipilimumab, nivolumab and combined, respectively. The uncured median survival times for the exponential model are 5 and 15 months, for PFS and OS, respectively.

The restricted mean survival time (RMST) is the average time-to-event over a fixed follow-up period and has been proposed as an intuitive alternative summary. For a follow-up time of 60 months and for the exponential model, the RMSTs (95% credible interval, CrI) for the uncured group are 20.6 (18.4, 23.1) and 6.66 (6.09, 7.39) months, for OS and PFS, respectively. For the OS end-point, the RMSTs are 28.7 (26.3, 31.1), 36.5 (34.1, 38.9) and 39.6 (37.1, 42.1) months, for ipilimumab, nivolumab and combined, respectively. For the PFS end-point, the RMSTs are 11.9 (10.0, 14.1), 24.3 (21.1, 27.4) and 27.7 (24.7, 30.8) months, for ipilimumab, nivolumab and combined, respectively. The RMST for the cured population is the same value corresponding to the general population of 58.2 months.

Figure 5 and Figure 6 show cure fraction posterior distribution forest plots for the hierarchical and separate mixture cure models, respectively. We can see that the global cure fraction posterior is relatively wide, which is understandable since there are only two exchangeable points providing limited information (OS and PFS). They approximately range 0.1-0.55, 0.2-0.6, 0.2-0.65 for ipilimumab, nivolumab and combined treatments respectively. However, despite this, the data influence the end-point cure fractions by pulling the central location away from the prior, downwards for ipilimumab and upwards for the combined treatment. For ipilimumab, the PFS curve fraction posterior is close to containing zero within the credible interval. Whereas for nivolumab and the combined treatment arms the mean global cure fraction lies somewhere between the PFS and OS means, the ipilimumab global mean is closer to the OS cure fraction mean. This is because we have explicitly codified in the prior distribution that the global cure fraction is unlikely to be near zero and the PFS cure fraction posterior is close to zero. We would observe the equivalent behaviour if the OS distributions were near to 1. The survival curves for the whole sample plateau to the background mortality survival from when the uncured survival probability reaches zero. Since the background is user-supplied, in this case from the WHO life tables, then there is no additional insight from this time on-wards. The times at which this occurs are relatively early for PFS and the log-Normal distribution. Comparing the hierarchical and separate models there is not a noticeable difference between the two. As mentioned above, there is limited information with only two end-points and a fairly weak prior. There is a small amount of shrinkage such that the posterior distributions for OS and PFS are pulled toward the global mean, so that the PFS cure fractions are larger and the OS cure fractions are smaller in the hierarchical model than the separate model.

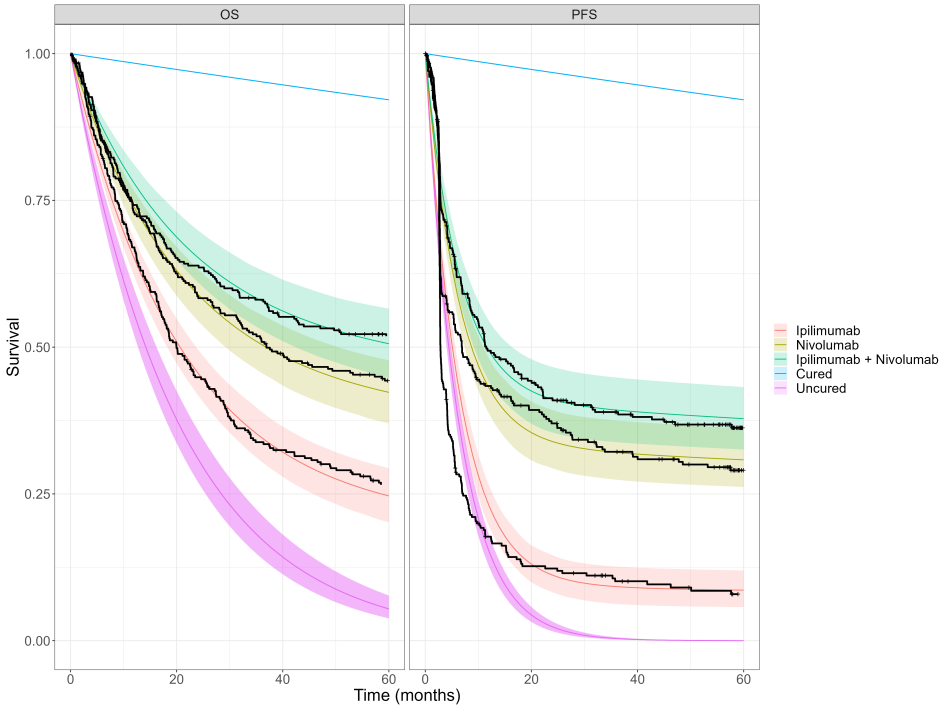

FIGURE 4 Bayesian hierarchical mixture cure model posterior mean survival curves for uncured fraction assumed exponential for both OS and PFS, and ipilimumab, nivolumab and combination treatments. The black lines show the Kaplan-Meier curves using the complete CheckMate 067 trial data.

4.6 | All data-cuts results

We now turn our attention to fitting the mixture cure models to earlier data-cuts in the CheckMate 067 trial data set. For the complete data set up to 60 months we have seen that the separate and hierarchical models produce similar parameter inferences. This is to be expected when the data are relatively mature, there are in effect only two data points at the global level and a weak prior is used for the global variance hyper parameter.

For demonstration purposes, we show results using the penalised complexity prior with exponential distribution rate parameter $\lambda = 55$ for the 12 months data cut scenario and retain all other previous priors, including the weak priors for the mean π_k. More data are available at later times and so the emphasis on this prior was relaxed. This approach is in the spirit of information-sharing methods such as power priors, which are "adaptively robust", where a power parameter controls the degree of borrowing. For example, for the simple case of normal historical data under a flat prior, the power prior is another normal distribution with a scaled variance according to the sample size and magnitude of the power parameter, determined by the investigator.

Figures 7 and Figure 8 show forest plots of the cure fraction estimates for 12, 30 and 60 months, and for the separate and hierarchical models, respectively. Results are shown for the OS exponential with PFS exponential model and the OS log-Normal with PFS log-Normal model.

We can see that in Figure 7 for the exponential model in the OS survival curves for the separate model there is more uncertainty than in the equivalent plot for the hierarchical model in Figure 8. The log-Normal model OS survival curves for separate and hierarchical models over-estimate the cure fraction at earlier data-cuts and all treatment arms. For later data-cuts, the cure fractions appear to converge from above to the complete data estimates.

Conversely, for the separate models and all treatment arms, at 12 months follow-up the exponential model underestimates the OS cure fraction and approaches the complete data value from below. This is less clear than for the log-Normal model probably because of the larger estimate uncertainty for the exponential model. Crucially, these are two clearly different behaviours but we would be ignorant to which is correct (if any) until later on in the trial. For OS in the hierarchical model, the cure fraction estimates appear to be accurate and more certain even at 12 months for all treatment arms. The posterior is more stable about the complete data cure fraction for all considered data-cuts.

TABLE 1 WAIC point estimates (and standard errors) for hierarchical model and all distributions.

OS distn	PFS distn	ELPD[†]	$p_D^{\ddagger}$	WAIC
exp	exp	-5064.82 (80.38)	11.16 (0.46)	10129.63 (160.77)
	Gompertz	-5065.64 (80.44)	11.8 (0.49)	10131.29 (160.88)
	loglogistic	-4970.34 (80.72)	11.78 (0.33)	9940.68 (161.45)
	lognormal	-4730.47 (92.79)	18.04 (0.91)	9460.94 (185.59)
	Weibull	-5062.99 (81.69)	14.63 (0.8)	10125.97 (163.38)
Gompertz	exp	-5064.99 (80.41)	11.45 (0.45)	10129.99 (160.82)
	Gompertz	-5065.61 (80.44)	11.66 (0.48)	10131.22 (160.88)
	loglogistic	-4970.42 (80.79)	11.68 (0.34)	9940.84 (161.59)
	lognormal	-4730.92 (92.64)	18.31 (0.95)	9461.84 (185.29)
	Weibull	-5063.81 (81.74)	15.45 (0.85)	10127.61 (163.48)
loglogistic	exp	-5059.86 (80.47)	11.29 (0.42)	10119.73 (160.93)
	Gompertz	-5060.37 (80.4)	11.64 (0.45)	10120.74 (160.8)
	loglogistic	-4966.22 (80.75)	12.77 (0.35)	9932.44 (161.5)
	lognormal	-4726.57 (93.17)	19.43 (0.91)	9453.13 (186.34)
	Weibull	-5058.17 (81.84)	14.8 (0.74)	10116.33 (163.67)
lognormal	exp	-4972.55 (84.58)	16.35 (0.61)	9945.09 (169.15)
	Gompertz	-4973.16 (84.69)	16.5 (0.56)	9946.32 (169.38)
	loglogistic	-4877.87 (85.79)	16.53 (0.5)	9755.74 (171.59)
	lognormal	-4638.85 (100.27)	23.58 (1)	9277.71 (200.55)
	Weibull	-4970.22 (86.17)	19.01 (0.89)	9940.44 (172.34)
Weibull	exp	-5063.26 (80.7)	11.93 (0.46)	10126.53 (161.4)
	Gompertz	-5064.37 (80.64)	12.67 (0.52)	10128.73 (161.28)
	loglogistic	-4968.74 (81.03)	12.3 (0.37)	9937.49 (162.07)
	lognormal	-4729.78 (93.18)	19.54 (0.97)	9459.56 (186.36)
	Weibull	-5061.91 (82.05)	15.77 (0.85)	10123.81 (164.11)

[†]ELPD: Expected log pointwise predictive density; [‡]p_D: Effective number of parameters; WAIC: widely applicable information criterion.

Figures 9 shows the survival curves for the data-cut with 12 months of follow-up data for the exponential and for the separate and hierarchical models, respectively. Equivalent Figures for 30 months data-cut are given in the Appendix.

Comparing the uncured RMSTs for various models, the hierarchical exponential model gives for OS 20.6 (18.4, 23.1) and PFS 6.72 (6.15, 7.34) months. For the separate model, OS is 22.3 (18.5, 26.9) and PFS is 5.68 (4.9, 6.45) months. For the log-Normal hierarchical model, the RMST for OS end-point is 6.6 (6.17, 7.13) and PFS is 3.98 (3.84, 4.12) months. Finally, the separate model gives 6.59 (6.18, 7.07) months for OS and 3.98 (3.85, 4.1) months for PFS.

Table 3 shows the survival estimates at 60 months extrapolated using 12 and 30 months data cut point. Clearly, when the survival curve has converged to the long-term plateau this is equivalent to the cure fraction.

5 | DISCUSSION

We have presented a Bayesian hierarchical mixture cure model to obtain complete survival curves which exhibit plateau behaviour. These curves can then be used in HTA to inform functions of patient lifetime. We have demonstrated the novel method using the CheckMate 067 trial data set for end-points OS and PFS and at different artificial data-cuts. The hierarchical model was compared with the separate model analogue and has been shown to be superior as well as both principled and applicable to situations where trials include multiple treatment arms and end-points.

A benefit to adopting a Bayesian paradigm over a frequentist approach is the flexibility to modelling complex structure, in our particular case multilevel or nested data. In contrast, a frequentist model would need to impose assumptions about the structure

TABLE 2 WAIC point estimates (and standard errors) for separate models and all distributions.

OS distn	PFS distn	ELPD†	$p_D^{\ddagger}$	WAIC
exp	exp	-5066.38 (81.4)	10.46 (0.43)	10132.76 (162.81)
	Gompertz	-5066.68 (81.46)	10.38 (0.46)	10133.37 (162.92)
	loglogistic	-4971.73 (81.67)	10.36 (0.28)	9943.46 (163.34)
	lognormal	-4732.41 (93.18)	18.56 (0.99)	9464.83 (186.36)
	Weibull	-5064.87 (82.86)	14.39 (0.86)	10129.74 (165.72)
Gompertz	exp	-5066.72 (81.44)	10.51 (0.45)	10133.43 (162.89)
	Gompertz	-5066.76 (81.49)	10.19 (0.44)	10133.53 (162.97)
	loglogistic	-4972.24 (81.7)	10.7 (0.28)	9944.47 (163.4)
	lognormal	-4732.23 (93.3)	18.01 (0.95)	9464.46 (186.61)
	Weibull	-5064.46 (82.92)	13.5 (0.78)	10128.91 (165.83)
loglogistic	exp	-5062.51 (81.89)	10.79 (0.4)	10125.02 (163.78)
	Gompertz	-5062.64 (82.01)	10.26 (0.41)	10125.27 (164.02)
	loglogistic	-4968.45 (82.36)	11.63 (0.28)	9936.89 (164.72)
	lognormal	-4728.25 (94.2)	18.18 (0.92)	9456.51 (188.4)
	Weibull	-5060.18 (83.31)	13.79 (0.69)	10120.37 (166.62)
lognormal	exp	-4973.43 (85.26)	14.63 (0.55)	9946.86 (170.53)
	Gompertz	-4973.48 (85.22)	14.31 (0.51)	9946.96 (170.44)
	loglogistic	-4879.07 (86.55)	15.07 (0.44)	9758.14 (173.09)
	lognormal	-4638.43 (100.4)	21.3 (0.96)	9276.85 (200.8)
	Weibull	-4972.48 (87.07)	19.54 (0.91)	9944.95 (174.13)
Weibull	exp	-5064.74 (81.52)	11.11 (0.45)	10129.48 (163.04)
	Gompertz	-5065.16 (81.59)	11.16 (0.44)	10130.32 (163.18)
	loglogistic	-4970.64 (81.87)	11.68 (0.35)	9941.27 (163.73)
	lognormal	-4730.93 (93.78)	19.04 (0.96)	9461.85 (187.56)
	Weibull	-5063.2 (83.09)	14.87 (0.83)	10126.41 (166.17)

†ELPD: Expected log pointwise predictive density; ‡p_D: Effective number of parameters; WAIC: widely applicable information criterion.

of a covariance matrix. Another advantage of adopting a hierarchical Bayesian approach is robustness to smaller sample sizes due to the provision of prior information and borrowing of strength between data for different end-points. Principled prior distribution specification about the cure fractions was shown to be especially useful early on in the trial when observable data about long-term survivors was weak. This results in the stabilisation of the cure fraction estimates. It may be that there is excessive borrowing of information between cure fractions when the end-points are distinctly different. In our example, this was not necessary, since there are only two end-points but for a larger number this may be appropriate. Partial-exchangeability can be used instead in this case[26].

In comparison to some previous frequentist approaches, our framework is able to include contextual information about the parameters of the model via the prior distributions. Incorporating sensible information, both in the form of external data sources or elicited from experts can stabilise the inference on the cure fraction and restrict/constrain its values via the prior distributions. For the analysis presented in this paper, we exploited different priors. Minimally informative priors allow the data to dominate the posterior even if the dataset is relatively small. In practice, there may be additional information from clinical experience or previous data that can better inform the prior selection. The priors can be specified on the natural scale to aid elicitation and interpretation. For example, we may have clinical knowledge that a drug treatment could generate a cure fraction greater than, say, 30%. If a (cancer) patient enters the study at 60 years old then there is a high degree of certainty that they will not live for, say, over 40 years. In the case of sparse data e.g. when there is a large amount of missing data perhaps due to right censoring, then 'regularising' the inference based on available prior information could make a significant and important difference. In particular, the method of blending curves is a recent, straightforward option for how to constrain survival[27]. How to elicit this prior information in practice may not be simple and a formal protocol should be adopted[28].

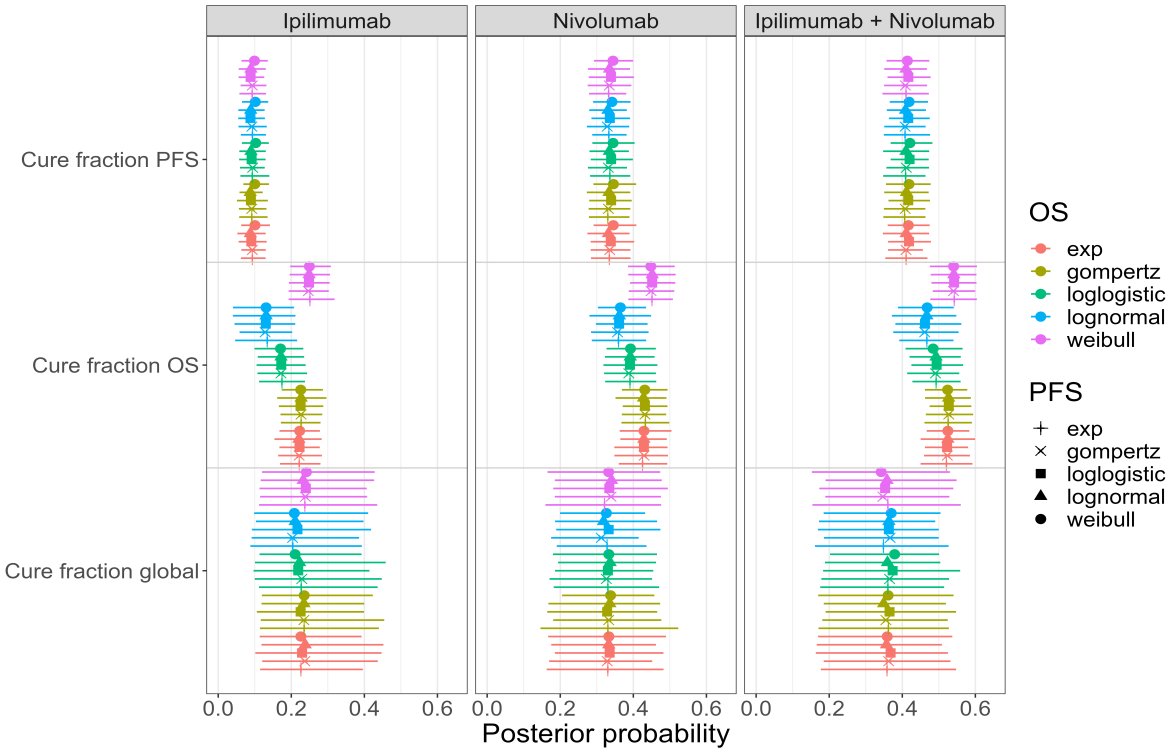

FIGURE 5 Hierarchical mixture cure model posterior cure fraction forest plots with 95% credible intervals for (i) ipilimumab only (ii) nivolumab only and (iii) combination treatment.

Healthcare regulators, such as the European Medicines Agency (EMA) and U.S. Food and Drug Administration (FDA), may value the fact that a hierarchical model can embed some form of prior knowledge to account for scepticism. This will prevent the cure rates to be taken at face value, particularly with limited data. The hierarchical model seems to give a more precise estimate of the cure rate, even with earlier data-cuts (especially for the exponential model). The hierarchical model is more aligned, which means the possibility of more reliable estimates early on, which is a good argument to complement modelling based on earlier cuts.

In our analysis, we simulated censoring due to data-cuts by simply censoring all patients in the complete data set at the same time point. In practice, a more realistic censoring time may have been at the date of last visit for each patient, but this would have made no differences for the purposes of this analysis. Primarily, the data-cut points were selected to demonstrate the theoretical justification of our method and not to strictly follow the study data-cut protocol.

Future work should test the proposed hierarchical MCM on a variety of data sets and at data-cuts reflecting their actual trial protocols. The modelling framework is generalisable from two end-points in our example to any number in principle. For example, previous related frequentist analyses considered clusters of hospitals and recurrent events per individual[10,9] or interspecies extrapolation in dose-response experiments[29]. Further work could investigate the benefit to including additional end-point data even if it does not clearly demonstrate plateauing behaviour. The background survival curve used in the model representing the statistically 'cured' survival in general population can be refined to be appropriate to specific cohort characteristics. For instance, long-term survival of clinically identified complete responders (CR) could be used to bridge the gap between statistical cure and clinical cure/functional cure. The shorter-term CR individual-level survival data can be combined with longer-term external data, such as from WHO life tables as in the current analysis (see[30] for potential approaches).

Another way to think about the cut-point data is to think about what is the minimum follow-up time should we choose to halt the trial and end the data collection which would maximise some expected overall utility. That is, put another way, what is the optimal stopping time τ^* for the total utility V?

$$\mathbb{E}V_{\tau^*} = \sup_{\tau} \mathbb{E}V_{\tau}$$

The stopping utility should incorporate a measure of the survival extrapolation robustness and may include benefits, such as smaller costs, freed-up time from the trial for patients and those running the study, quicker time to market for a drug, and

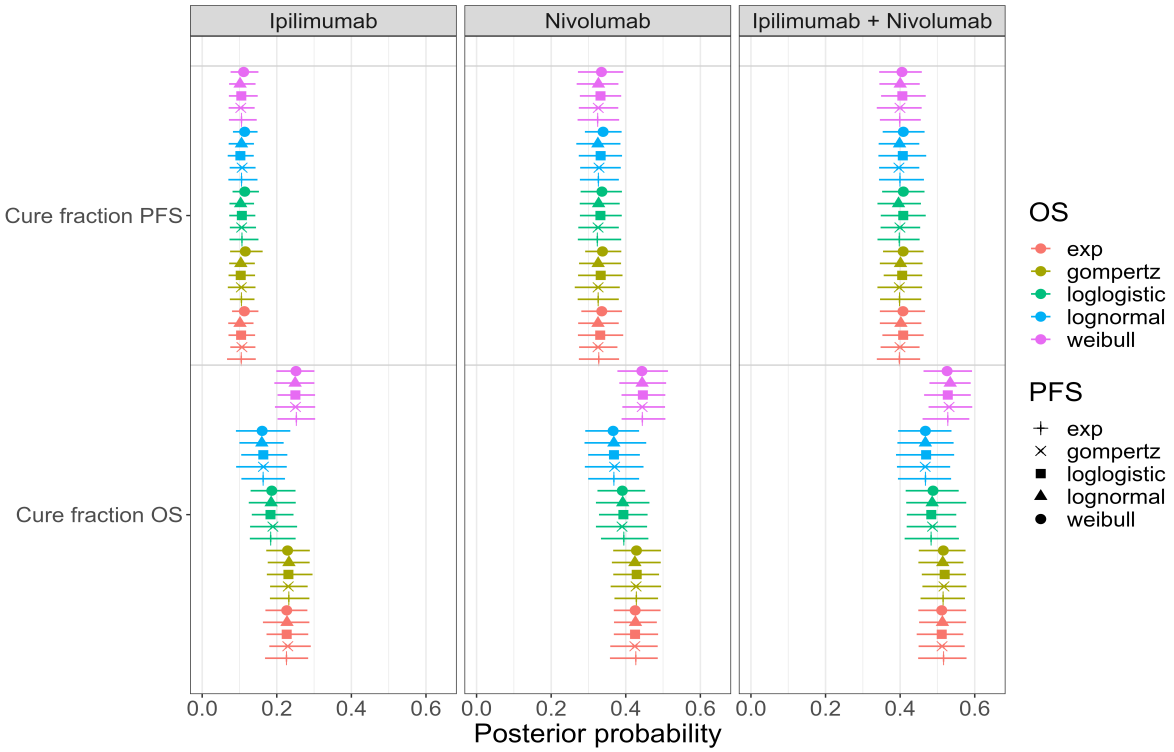

FIGURE 6 Separate mixture cure models posterior cure fraction forest plots with 95% credible intervals for (i) ipilimumab only (ii) nivolumab only and (iii) combination treatment.

avoidance of side-effects etc. The utility would also depend on estimating 'correct' parameters values from the available data ie, that they are within some acceptable threshold of the true values. This could be applied directly on the parameters of interest or an order statistics for which drug is deemed preferable.

This may be framed in terms of value of information (VoI)[31]. For example, in the case given above, what is the additional value of obtaining further data after 12 and 30 months. In practice, deriving a set of heuristics may be better than formalising a general rule. A discrete set of cut-points may be used, similar to our artificial selection of rounded times 12 and 30 months, rather than considering the whole real line.

As well as during a study, the hierarchical MCM could be used in advance of a study to inform trial design and optimal follow-up time to support the choice of data-cuts used in practice. Smaller sample sizes and shorter duration may provide equivalent power given the exploitation of additional model structure.

There are several implications for HTA by using the hierarchical mixture cure model presented here. In a cost-effectiveness analysis, unstable estimates over time would give different estimates of benefit and thus cost-effectiveness statistics and potentially differing optimal decisions. Lee (2019)[32] performed a cost-effectiveness analysis of using nivolumab with ipilimumab vs ipilimumab using the CheckMate 067 trial data, and a partitioned survival model and Markov state-transition model. Lifetime costs and benefits were estimated. Results were obtained using 18-month (when OS data were unavailable) and 36-month (OS available) CheckMate 067 data-cuts. They showed that the model using OS data generated more than 1 additional quality-adjusted life-year (QALY) across both treatment arms compared to the model without OS data. Using our approach, we have shown that reliable estimates of complete survival curves for OS can be obtained even for early data-cuts. These can be used to estimate lifetime horizon benefits in instances where perhaps previously estimates were not available. In our example, the OS cure fraction is accurate and stable at 12 months for the hierarchical mixture cure model with exponential distributions for the OS and PFS uncured survival models. Additional benefits are potential early end of trial (cheaper, improved health, quicker drugs to market for producers and patients benefit), and smaller trials.

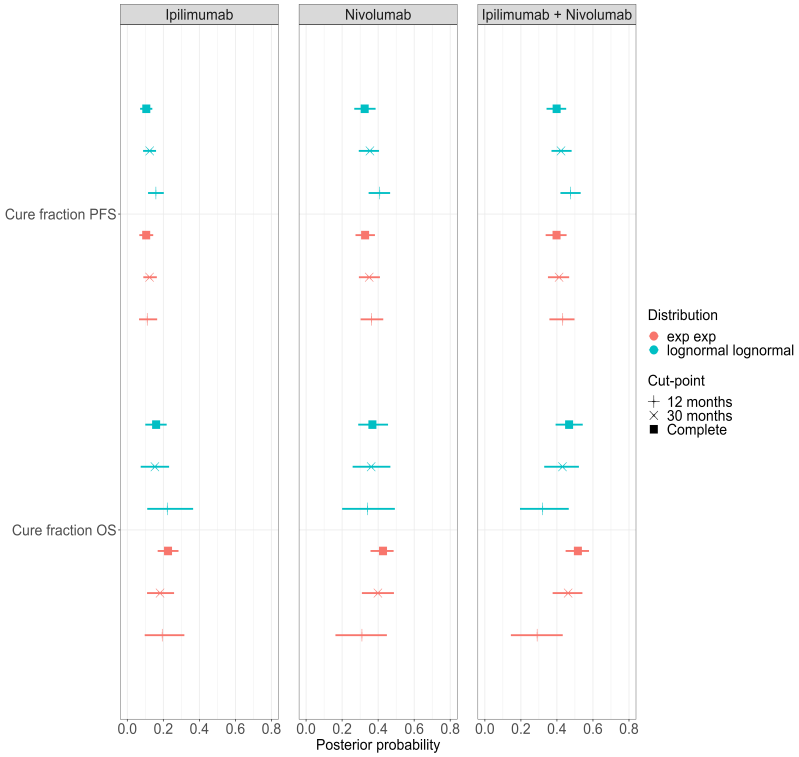

FIGURE 7 Posterior cure fraction forest plot with 95% credible intervals for study data with cutpoint censoring at 12 months, 30 months and the complete data set using the separate mixture cure models.

References

1. Latimer N. NICE DSU Technical Support Document No. 14: Survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. *London: National Institute for Health and Care Excellence (NICE)* 2013.
2. Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Statistics in Medicine* 2006; 25(11): 1960–1975. doi: 10.1002/sim.2366
3. Jackson CH, Sharples LD, Thompson SG. Survival models in health economic evaluations: balancing fit and parsimony to improve prediction.. *The international journal of biostatistics* 2010; 6(1): Article 34. doi: 10.2202/1557-4679.1269
4. Ouwens MJ, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs AH. Estimating Lifetime Benefits Associated with Immuno-Oncology Therapies: Challenges and Approaches for Overall Survival Extrapolations. *Pharmacoeconomics* 2019; 37(9): 1129–1138. doi: 10.1007/s40273-019-00806-4
5. Khair DO, Bax HJ, Mele S, et al. Combining immune checkpoint inhibitors: Established and emerging targets and strategies to improve outcomes in melanoma. *Front. Immunol.* 2019; 10(MAR): 1–20. doi: 10.3389/fimmu.2019.00453
6. Wolchok JD, Chiarion-Sileni V, Gonzalez R, et al. Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N. Engl. J. Med.* 2017; 377(14): 1345–1356. doi: 10.1056/NEJMoa1709684
7. Larkin J, Chiarion-Sileni V, Gonzalez R, et al. Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *New England Journal of Medicine* 2019; 381(16): 1535–1546. doi: 10.1056/nejmoa1910836
8. Bullement A, Willis A, Amin A, Schlichting M, Hatswell AJ, Bharmal M. Evaluation of survival extrapolation in immuno-oncology using multiple pre-planned data cuts: Learnings to aid in model selection. *BMC Med. Res. Methodol.* 2020; 20(1): 1–14. doi: 10.1186/s12874-020-00997-x

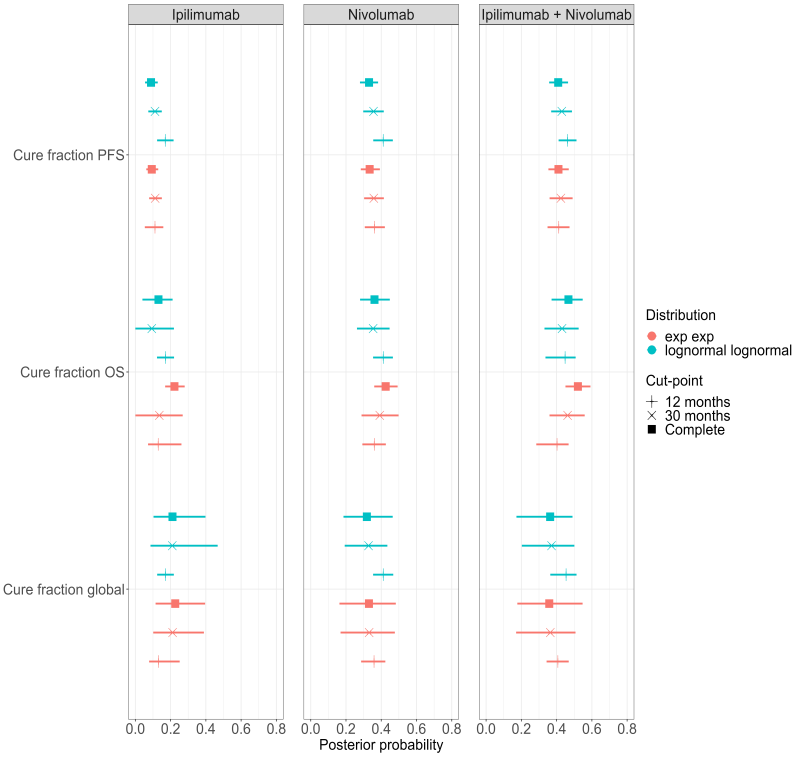

FIGURE 8 Posterior cure fraction forest plot with 95% credible intervals for study data with cutpoint censoring at 12 months, 30 months and the complete data set using the hierarchical mixture cure model.

9. Lai X, Yau KKW. Multilevel Mixture Cure Models with Random Effects. *Biometrical Journal* 2009; 51(3): 456–466. doi: 10.1002/bimj.200800222

10. Lai X, Yau KKW. Long-term survivor model with bivariate random effects: Applications to bone marrow transplant and carcinoma study data. *Statistics in Medicine* 2008; 27(27): 5692–5708. doi: 10.1002/sim.3404

11. Tan SH, Abrams KR, Bujkiewicz S. Bayesian Multiparameter Evidence Synthesis to Inform Decision Making: A Case Study in Metastatic Hormone-Refractory Prostate Cancer. *Medical Decision Making* 2018; 38(7): 834–848. doi: 10.1177/0272989X18788537

12. Hodi FS, Chiarion-Sileni V, Gonzalez R, et al. Nivolumab plus ipilimumab or nivolumab alone versus ipilimumab alone in advanced melanoma (CheckMate 067): 4-year outcomes of a multicentre, randomised, phase 3 trial. *Lancet Oncol.* 2018; 19(11): 1480–1492. doi: 10.1016/S1470-2045(18)30700-9

13. Hodi F, Chiarion-Sileni V, Lewis K, et al. Long-term survival in advanced melanoma for patients treated with nivolumab plus ipilimumab in CheckMate 067. *Journal of Clinical Oncology* 2022; 40: 9522-9522. doi: 10.1200/JCO.2022.40.16$_{s}$*uppl*.9522

14. Mohr P, Larkin J, Paly V, et al. 1105P Estimating long-term survivorship in patients with advanced melanoma treated with immune-checkpoint inhibitors: Analyses from the phase III CheckMate 067 trial. *Annals of Oncology* 2020; 31: S747. doi: 10.1016/J.ANNONC.2020.08.1228

15. Yu X, De Angelis R, Andersson T, Lambert P, O'Connell D, Dickman P. Estimating the proportion cured of cancer: Some practical advice for users. *Cancer Epidemiology* 2013; 37(6): 836–842. doi: 10.1016/j.canep.2013.08.014

16. Nikolaidis GF, Woods B, Palmer S, Soares MO. Classifying information-sharing methods. 2021: 1–13.

17. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. *Journal of statistical software* 2017; 76(1).

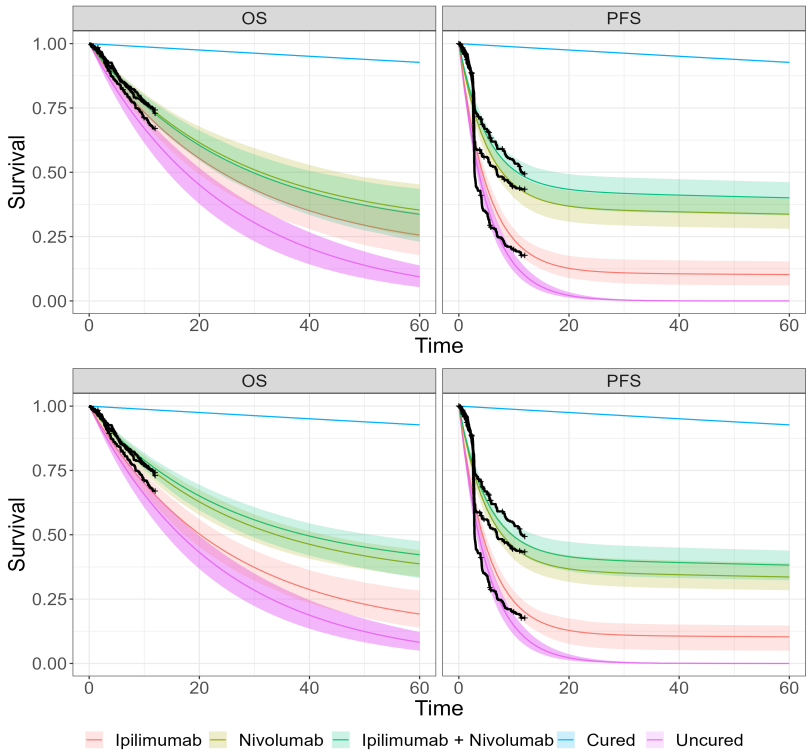

FIGURE 9 Mixture cure models posterior survival curves with 95% credible intervals using cut-point data for 12 month follow-up and for separate (top) and hierarchical (bottom) models. Both PFS and OS use exponential uncured survival curves. The black lines show the observed data Kaplan-Meier curves.

18. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2020.

19. WHO . Life tables by country. 2020. http://apps.who.int/gho/data/ (accessed on 02/12/20).

20. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; 1(3): 515-–533.

21. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* 2013(4): 685-–709.

22. Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* 2017; 32(1): 1–28. doi: 10.1214/16-STS576

23. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 2017; 27(5): 1413–1432. doi: 10.1007/s11222-016-9696-4

24. Wileyto EP, Li Y, Chen J, Heitjan DF. Assessing the fit of parametric cure models. *Biostatistics* 2013; 14(2): 340–350. doi: 10.1093/biostatistics/kxs043

25. Peng Y, Yu B. *Cure Models*. Chapman and Hall/CRC . 2021

26. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics* 2016; 15(2): 123–134. doi: 10.1002/pst.1730

27. Che Z, Green N, Baio G. Blended Survival Curves: A New Approach to Extrapolation for Time-to-Event Outcomes from Clinical Trials in Health Technology Assessment. *Medical Decision Making* 2022; 25(1): 0272989X2211345. doi: 10.1177/0272989X221134545

TABLE 3 Survival probability estimates for separate and hierarchical models at 60 months using 12, 30 and 60 months data cut points. Models with exponential or log-normal distributions for both OS and PFS are given.

Endpoint	Model	Distribution	Cut-point (month)	Ipilimumab	Nivolumab	Combined
OS	hier	exp	12	0.19 [0.14, 0.28]	0.34 [0.28, 0.39]	0.42 [0.33, 0.48]
OS	hier	exp	30	0.2 [0.11, 0.28]	0.33 [0.28, 0.38]	0.47 [0.4, 0.55]
OS	hier	exp	60	0.25 [0.2, 0.29]	0.31 [0.26, 0.36]	0.51 [0.46, 0.56]
OS	hier	lognormal	12	0.28 [0.23, 0.33]	0.38 [0.33, 0.43]	0.49 [0.43, 0.54]
OS	hier	lognormal	30	0.23 [0.17, 0.3]	0.33 [0.28, 0.38]	0.49 [0.43, 0.56]
OS	hier	lognormal	60	0.25 [0.21, 0.3]	0.31 [0.26, 0.35]	0.51 [0.46, 0.56]
OS	separate	exp	12	0.26 [0.18, 0.34]	0.34 [0.28, 0.4]	0.34 [0.23, 0.44]
OS	separate	exp	30	0.22 [0.17, 0.28]	0.32 [0.27, 0.38]	0.47 [0.4, 0.52]
OS	separate	exp	60	0.25 [0.21, 0.3]	0.3 [0.25, 0.35]	0.5 [0.45, 0.55]
OS	separate	lognormal	12	0.34 [0.26, 0.42]	0.38 [0.32, 0.43]	0.41 [0.33, 0.51]
OS	separate	lognormal	30	0.27 [0.22, 0.32]	0.33 [0.27, 0.37]	0.48 [0.42, 0.54]
OS	separate	lognormal	60	0.27 [0.22, 0.32]	0.3 [0.25, 0.35]	0.51 [0.45, 0.56]
PFS	hier	exp	12	0.1 [0.05, 0.15]	0.39 [0.34, 0.44]	0.38 [0.32, 0.44]
PFS	hier	exp	30	0.1 [0.07, 0.14]	0.41 [0.34, 0.49]	0.39 [0.33, 0.45]
PFS	hier	exp	60	0.08 [0.05, 0.12]	0.42 [0.37, 0.48]	0.38 [0.32, 0.44]
PFS	hier	lognormal	12	0.16 [0.11, 0.2]	0.47 [0.41, 0.51]	0.43 [0.38, 0.48]
PFS	hier	lognormal	30	0.1 [0.07, 0.14]	0.43 [0.38, 0.49]	0.4 [0.34, 0.45]
PFS	hier	lognormal	60	0.08 [0.05, 0.12]	0.43 [0.38, 0.48]	0.38 [0.32, 0.43]
PFS	separate	exp	12	0.1 [0.06, 0.15]	0.35 [0.25, 0.45]	0.4 [0.33, 0.46]
PFS	separate	exp	30	0.11 [0.08, 0.15]	0.41 [0.34, 0.48]	0.38 [0.32, 0.43]
PFS	separate	exp	60	0.1 [0.07, 0.14]	0.42 [0.37, 0.48]	0.37 [0.31, 0.42]
PFS	separate	lognormal	12	0.15 [0.11, 0.19]	0.43 [0.35, 0.52]	0.44 [0.39, 0.49]
PFS	separate	lognormal	30	0.12 [0.08, 0.15]	0.43 [0.36, 0.5]	0.39 [0.34, 0.45]
PFS	separate	lognormal	60	0.1 [0.07, 0.13]	0.43 [0.38, 0.48]	0.37 [0.31, 0.42]

28. O'Hagan A. Expert Knowledge Elicitation: Subjective but Scientific. *American Statistician* 2019; 73(sup1): 69–81. doi: 10.1080/00031305.2018.1518265

29. DuMouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association* 1983; 78(382): 293–308. doi: 10.1080/01621459.1983.10477968

30. Jackson C, Stevens J, Ren S, et al. Extrapolating Survival from Randomized Trials Using External Data: A Review of Methods. *Medical Decision Making* 2017; 37(4): 377–390. doi: 10.1177/0272989X16639900

31. Heath A, Manolopoulou I, Baio G. A Review of Methods for Analysis of the Expected Value of Information. *Medical Decision Making* 2017; 37(7): 747–758. doi: 10.1177/0272989X17697692

32. Lee D, Amadi A, Sabater J, et al. Can We Accurately Predict Cost Effectiveness Without Access to Overall Survival Data? The Case Study of Nivolumab in Combination with Ipilimumab for the Treatment of Patients with Advanced Melanoma in England. *PharmacoEconomics - Open* 2019; 3(1): 43–54. doi: 10.1007/s41669-018-0080-5

How to cite this article: Green N. et al (2024), , *1; 00:1–6*.