



This is a repository copy of *Assurance methods for designing a clinical trial with a delayed treatment effect*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/212988/>

Version: Published Version

Article:

Salsbury, J.A., Oakley, J.E. orcid.org/0000-0002-9860-4093, Julious, S.A. et al. (1 more author) (2024) Assurance methods for designing a clinical trial with a delayed treatment effect. *Statistics in Medicine*, 43 (19). pp. 3595-3612. ISSN 0277-6715

<https://doi.org/10.1002/sim.10136>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Assurance methods for designing a clinical trial with a delayed treatment effect

James A. Salsbury¹  | Jeremy E. Oakley¹ | Steven A. Julious²  | Lisa V. Hampson³

¹The School of Mathematics and Statistics, The University of Sheffield, Sheffield, UK

²The School of Health and Related Research, The University of Sheffield, Sheffield, UK

³Advanced Methodology and Data Science, Novartis Pharma AG, Basel, Switzerland

Correspondence

James A. Salsbury, The School of Mathematics and Statistics, The Hicks Building, Broomhall, Sheffield, S3 7RH, UK.

Email: jsalsbury1@sheffield.ac.uk

Funding information

Engineering and Physical Sciences Research Council, Grant/Award Number: 2610753

An assurance calculation is a Bayesian alternative to a power calculation. One may be performed to aid the planning of a clinical trial, specifically setting the sample size or to support decisions about whether or not to perform a study. Immuno-oncology is a rapidly evolving area in the development of anticancer drugs. A common phenomenon that arises in trials of such drugs is one of delayed treatment effects, that is, there is a delay in the separation of the survival curves. To calculate assurance for a trial in which a delayed treatment effect is likely to be present, uncertainty about key parameters needs to be considered. If uncertainty is not considered, the number of patients recruited may not be enough to ensure we have adequate statistical power to detect a clinically relevant treatment effect and the risk of an unsuccessful trial is increased. We present a new elicitation technique for when a delayed treatment effect is likely and show how to compute assurance using these elicited prior distributions. We provide an example to illustrate how this can be used in practice and develop open-source software to implement our methods. Our methodology has the potential to improve the success rate and efficiency of Phase III trials in immuno-oncology and for other treatments where a delayed treatment effect is expected to occur.

KEYWORDS

assurance, delayed treatment effects, expert judgement, prior elicitation, probability of success

1 | INTRODUCTION

Assurance calculations are growing in popularity as an aid for the design of clinical trials. An assurance calculation is a Bayesian alternative to a power calculation: instead of assuming parameters (eg, related to treatment effects) take particular values, we elicit prior distributions for them, enabling us to derive a probability of a successful trial outcome, accounting for uncertainty about treatment effects. The concept of an assurance calculation was first considered by Spiegelhalter and Freedman,¹ then developed by O'Hagan et al,² who coined the term 'assurance'. Note that the assurance method has had other terms accredited to it, such as average power, expected power and predictive power.³

To calculate assurance, we sample from the elicited prior distributions for the unknown parameters and then simulate clinical trials using these sampled values. The prior predictive probability that the trial will be 'successful' is the

proportion of simulated trials that meet our stated success criteria. These success criteria are not fixed by the assurance method; instead, they are set independently by the sponsor and can be any criteria that the sponsor wishes to consider (eg, that the observed treatment effect will be positive; or statistically significant; or exceed a clinically relevant threshold). More recently, assurance has been used to calculate the probability of obtaining regulatory approval with clinically relevant effects on key endpoints after Phase IIb.^{4,5} Note that the method for the trial data analysis is also specified independently of the assurance method; the same analysis method would be assumed as that used in the power calculation.

Assurance provides a more realistic assessment of the probability a trial will give rise to a successful outcome compared to a conventional power calculation. The high failure rates of clinical trials are well-documented,⁶ and high expectations based on Phase II data may not always translate to pivotal trial success.⁷ Assurance calculations for Phase III trials should capture the strength of the available evidence after mitigating the selection bias often inherent in early phase data when a necessary condition for progress is positive Phase Ib or Phase II results. They should also account for any limitations in the available data in light of planned shifts in the patient population, outcome or treatment strategies between phases.

Accurate and reliable evaluations of risk can be used to optimize trial design and analysis plans. For example, assurance can be used to support decisions regarding study sample size, and quantitatively measure how effective various trial setups are at reducing risk, such as the timing and number of planned interim analyses.^{8,9} Furthermore, assurance evaluations can also enable better informed decisions on whether or not to conduct a study. Of course sponsors, such as pharmaceutical companies and public funding bodies, may choose to fund a Phase III clinical trial (or indeed a program of Phase III clinical trials) regardless of whether it has a low assurance if the corresponding expected net present value (eNPV) is sufficiently high, thus targeting resources towards research programs with the greatest expected impact for patients.

Immuno-oncology (IO) is a rapidly evolving area in the development of anticancer drugs. In trials of IO therapies, time-varying treatment effects that deviate from the proportional hazards (PH) assumption have been observed on time-to-event endpoints such as progression-free survival (PFS) and overall survival (OS). See for example, Check-Mate 017.¹⁰ In a systematic review of 63 confirmatory randomized controlled trials (RCTs) of anti-programmed cell death protein-1 and anti-programmed death/ligand 1 therapies,¹¹ 15 instances were identified of suspected nonproportional hazards due to reasons including crossing of the OS survival curves^{12,13} or a lag before the PFS survival curves separated.¹⁴ In what follows, we focus on the latter scenario and refer to this as a delayed treatment effect (DTE).

There are several challenges associated with the design and analysis of trials with nonproportional hazards. Firstly, the primary estimand should be defined with a clinically interpretable measure used to summarize the benefit of the test treatment versus control,¹⁵ and an unbiased estimator should be selected to target it. Secondly, the test of the null hypothesis of no benefit of treatment versus control should be carefully selected acknowledging the impact of potential deviations from PH on the attained power of commonly applied procedures, such as the log-rank test.¹⁶ Where we suspect (but are not certain) that there will be a delay in the treatment effect, and furthermore are uncertain about the length of the delay if there is one, the target event number and corresponding sample size needs to be carefully chosen to provide confidence the trial will be able to meet its objectives in light of these uncertainties.

As IO trials are becoming more common, so are trials in which a DTE is observed. However, to the best of our knowledge, there has been no published work on eliciting prior distributions and calculating assurance for when a DTE is likely to be present in a clinical trial with time-to-event endpoints. In this article, we propose a method for how to elicit the relevant parameters for this trial and how to perform an assurance calculation.

In Section 2, we briefly discuss the assurance method and how it is used in practice. In Section 3, we define DTEs, present an elicitation method and signpost to the open-source software we have developed for use in this situation. In Section 4, we illustrate how our method can be used to calculate assurance. In Section 5, we investigate the robustness of our parameterisation and lastly we conclude with a brief summary in Section 6.

2 | ASSURANCE

Suppose that an RCT is to be conducted to compare an experimental treatment with a control; we assume that this is the current standard of care, but could also be a placebo. We want to test the null hypothesis H_0 that the treatment effect $\theta = 0$ versus the alternative hypothesis H_1 that $\theta \neq 0$. For a power calculation, the sample size is chosen to solve

$$P(\text{Reject } H_0 | \theta = \theta_A) = \pi^*, \quad (1)$$

for some desired probability π^* (usually 80% or 90%) and treatment effect θ_A , typically chosen to represent a plausible and clinically relevant effect.

The power of the test of H_0 at θ_A is the probability of rejecting H_0 if θ is as large as θ_A . However, since the θ may differ from θ_A , the attained power of the test may deviate from the target π^* . Assurance is the unconditional probability that the trial will end with the desired outcome:

$$P(\text{'Successful trial'}) = \int P(\text{'Successful trial'} | \theta) f(\theta) d\theta, \quad (2)$$

where $f(\theta)$ is the prior distribution for θ . If a successful trial simply corresponds to rejecting H_0 , Equation (2) is the *expected power*, interpreting θ_A in Equation (1) as the true value of the treatment effect, rather than some minimum clinically relevant difference.

If the desired outcome is to reject H_0 with data which favours the experimental treatment, then the event 'successful trial' may be defined as 'Reject H_0 with $\hat{\theta} > 0$ '. When calculating assurance, a key question is how to define the prior distribution $f(\theta)$ for the unknown treatment effect. One approach would be to take $f(\theta)$ as the posterior distribution for θ resulting from using clinical data from an early Phase II trial to update a weakly informative prior distribution. However, this approach may fail to incorporate other sources of relevant information, and may become challenging if there are differences between the treatment effect studied in Phase II and the quantity of interest in the future trial. Alternatively, the prior distribution(s) for the parameters of interest could be elicited from a group of experts, in light of the Phase IIb trial data and any other information that is deemed relevant—data from drugs with a similar mechanism of action, knowledge about the disease area, and so forth. For a detailed discussion about the method of eliciting parameters in these contexts, see Dallow et al.⁸

The reason expert elicitation is useful in these circumstances is to bridge the gap between data from the completed Phase IIb trial and the quantities of interest in the planned Phase III trial. For example, the future trial may consider different endpoints, the patient population may change, or a different dose/dosing regime may be proposed.¹⁷ Also, when working in a rare disease setting, there may be limited data available. In this context, expert elicitation is useful as it allows the study team to combine heterogeneous sources of information (RCTs, case series, observational data) when a formal mathematical synthesis of these data would be very complex.¹⁸

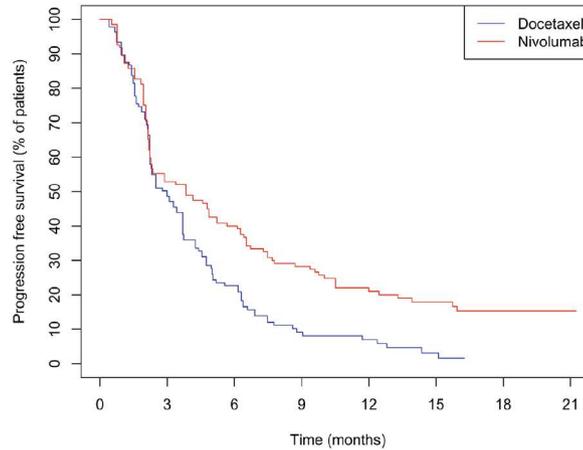
In the context of assurance methods, O'Hagan et al.² considered eliciting beliefs for clinical trials with Normally distributed and dichotomous endpoints. Gasparini et al.¹⁹ also considered Normally distributed endpoints, and Alhussain and Oakley²⁰ considered eliciting uncertainty about the variance of Normally distributed endpoints. For time-to-event outcomes, Spiegelhalter et al.²¹ considered stipulating a Normal prior distribution for the log hazard ratio under a PH assumption, as did Hiance et al.²² Ren and Oakley²³ considered expert elicitation for both parametric and non-parametric models. Azzolina et al.²⁴ produced a comprehensive literature review of assurance methods that use expert elicitation (both theoretical and applied).

3 | METHODS

3.1 | Delayed treatment effects

Figure 1 shows a Kaplan-Meier plot from a Phase III trial, CheckMate 017,¹⁰ in which a DTE was observed. The trial enrolled patients with advanced squamous-cell non-small-cell lung cancer (NSCLC) and compared the current standard of care, docetaxel, against an experimental treatment, nivolumab. The plot is based on the reconstructed individual patient data²⁵ derived from published Kaplan-Meier survival curves. We see that both the control and experimental treatment curves follow the same trajectory for some time (approximately 3 months), after which they separate.

In a survival trial, suppose we have two groups: the control group and the experimental treatment group. We denote the hazard function for the control group as $h_c(t)$ and the hazard function for the experimental treatment group as $h_e(t)$. In a typical survival trial, H_1 assumes that the hazard function for the experimental treatment group is less than or equal to the hazard function for the control group at all time points, that is $H_1: h_e(t) \leq h_c(t), \forall t$. This suggests that patients in the experimental treatment arm immediately benefit from the intervention compared to those in the control arm.



No. at Risk	
Nivolumab	135 68 48 33 21 15 6 2
Docetaxel	137 62 26 9 6 2 1 0

FIGURE 1 Kaplan-Meier plot of a Phase III trial, Checkmate 017,¹⁰ in which DTEs are present. The control and experimental treatment curves follow the same trajectory for approximately 3 months, after which they separate.

In a trial in which a DTE is thought likely to occur, we make a different assumption. We assume that the hazard function for the experimental treatment group is the same as that of the control group until a certain time T , which represents the delay in the experimental treatment taking effect. After time T , we assume that the experimental treatment group starts experiencing some benefit relative to the control group:

$$h_e(t) = \begin{cases} h_c(t), & t \leq T \\ h_e^*(t), & t > T \end{cases} \tag{3}$$

where $h_e^*(t) \leq h_c(t)$ describes the benefit of the experimental treatment relative to control.²⁶

In survival trials, the PH assumption is often made. It is used in Cox regression and the log-rank test, which is a standard statistical test in survival trials, is most powerful under this assumption. However, when DTEs are present, this assumption is violated because the hazard ratio becomes time-dependent. This poses challenges for the design and analysis of trials with DTEs.

Various researchers have proposed methodologies to address trials with a DTE.^{26–36} However, most of these discussions focus on regaining statistical power lost due to the delay by using alternative analysis methods, such as weighted log-rank tests or the difference in restricted mean survival times (RMST).³⁷ These methods aim to account for the time-dependent hazard ratio without assuming PH or the specific shape of the underlying survival curves.

3.2 | Assurance for delayed treatment effects

We propose an elicitation technique and parameterisation to calculate assurance in these circumstances. By doing so, we are able to capture experts' uncertainty about the relevant parameters and provide a more realistic judgement of the probability of success of the proposed trial.

We suppose that the survival times in the control group follow a Weibull distribution with hazard function

$$h_c(t) = \gamma_c \lambda_c^{\gamma_c} t^{\gamma_c - 1} \tag{4}$$

and corresponding survival function

$$S_c(t) = \exp\{-(\lambda_c t)^{\gamma_c}\}. \tag{5}$$

We assume survival times in the experimental treatment group, after a delay of length T , also follow a Weibull distribution with different parameters to the control. This induces the hazard function

$$h_e^*(t) = \gamma_e \lambda_e^{\gamma_e} t^{\gamma_e - 1} \quad (6)$$

and corresponding survival function

$$S_e^*(t) = \exp\{-(\lambda_c T)^{\gamma_c} - \lambda_e^{\gamma_e} (t^{\gamma_e} - T^{\gamma_e})\}. \quad (7)$$

Prior to time T , we assume the experimental treatment group has the same survival function as the control (Equation 5). Thus, the survival function for the experimental treatment group is

$$S_e(t) = \begin{cases} \exp\{-(\lambda_c t)^{\gamma_c}\}, & t \leq T \\ \exp\{-(\lambda_c T)^{\gamma_c} - \lambda_e^{\gamma_e} (t^{\gamma_e} - T^{\gamma_e})\}, & t > T \end{cases} \quad (8)$$

The hazard ratio of the two groups (derived from Equations 4 and 6) is

$$\text{HR}(t) = \begin{cases} 1, & t \leq T \\ \frac{\gamma_e \lambda_e^{\gamma_e} t^{\gamma_e - 1}}{\gamma_c \lambda_c^{\gamma_c} T^{\gamma_c - 1}}, & t > T \end{cases} \quad (9)$$

3.3 | Constructing the prior distributions

From Equations (5) and (8), we see that there are five unknowns: T , λ_c , γ_c , λ_e and γ_e . To calculate assurance, prior distributions are required for these parameters. In the following sections we propose a method for eliciting these priors, including the questions to ask.

3.3.1 | Prior(s) for λ_c and γ_c

We first elicit judgements on the two parameters for the survival times in the control group; λ_c , also known as the scale parameter, and γ_c , the shape parameter. We assume that there exists some historical data on the control so that we can derive

$$\pi(\lambda_c, \gamma_c | \mathbf{x}_{\text{hist}}),$$

where \mathbf{x}_{hist} is the historical data for the control group intervention. Schmidli et al³⁸ consider using a meta-analytic-predictive (MAP) prior for control group parameters. Bertsche et al³⁹ extend this method to specifically consider time-to-event data. Alternatively, to include expert elicitation at this stage, see Ren and Oakley,²³ who consider eliciting beliefs when survival times are assumed to follow a Weibull distribution.

3.3.2 | Prior for T

We propose a hierarchical procedure for eliciting judgements about T , λ_e and γ_e , as shown in Figure 2. The existence of a DTE presupposes that the treatment has any effect in the first place, but the experts may not be certain of this. Hence we first need to elicit a probability that the treatment has any effect, and then elicit judgements about T conditional on the assumption that the treatment has some effect. To avoid ambiguity, we define “a treatment effect” as any separation between the survival curves for the control and treatment groups: they are not equal. Hence the first question we ask is

“What is your probability that the population survival curves separate at some point in time?”

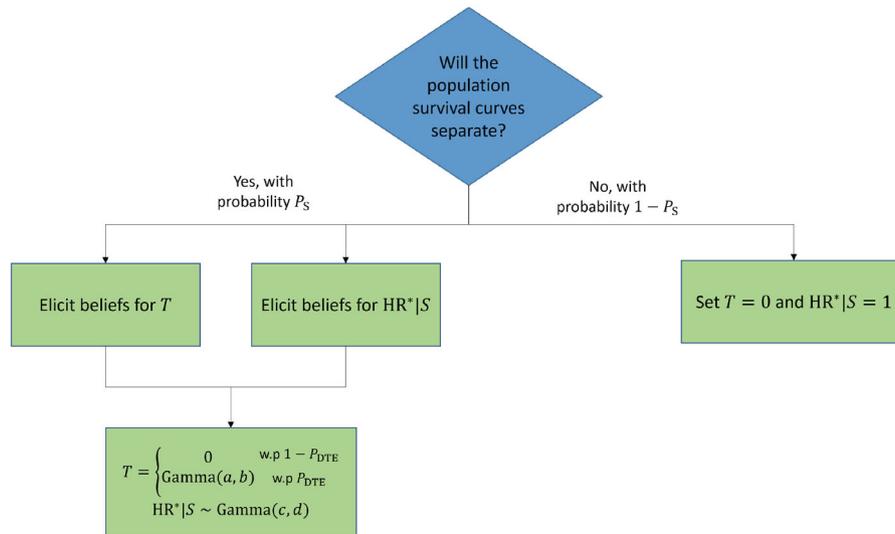


FIGURE 2 The proposed elicitation scheme as described in Sections 3.3.2 and 3.3.3.

We define S to be the proposition that the population survival curves separate, and P_S to be elicited probability that this proposition is true. Note that the proposition S refers to the unobserved true population distributions of survival, and not sampled Kaplan-Meier curves that would be observed in a trial. We now elicit judgements about T , conditional on S . Given S , we allow for the possibility of no delay in the treatment effect; we propose a prior of the form

$$T = \begin{cases} 0, & \text{with probability } P_{DTE} \\ D_{\text{delay}}, & \text{with probability } 1 - P_{DTE} \end{cases}, \tag{10}$$

with $D_{\text{delay}} \sim \text{Gamma}(a, b)$. Any non-negative distribution could be used for D_{delay} but we expect the Gamma distribution to be sufficiently flexible. We therefore need questions that an expert would be willing to answer and from which we can identify values for P_{DTE} , a and b . We elicit judgements about the probability P_{DTE} by asking the following question:

“If we suppose that the population survival curves separate at some time, what is your probability that there is a delay before they separate?”

Finally, we elicit judgements about the distribution of D_{delay} by stating

“Suppose that the population survival curves separate with a delay. We want you to consider your uncertainty about the delay.”

We then use a standard method for eliciting a univariate distribution for D_{delay} . Methods for eliciting univariate distributions can be found in O’Hagan et al⁴⁰ and implemented using the Sheffield Elicitation Framework (SHELF).⁴¹ SHELF is a package of protocols, templates and guidance documents for conducting expert elicitation. There are various methods that SHELF uses to elicit distributions that involve asking an expert to provide quantile judgements (eg, a median; tertiles) or probability judgements (the probability of the uncertain quantity lying in some interval). In either case, the expert is, in effect, specifying points on their cumulative distribution function. Parametric distributions are fitted to these judgements using a least squares procedure: the parameters are chosen to ensure the points on the fitted cumulative distribution function are as close as possible to the elicited. Feedback (additional quantiles or probabilities from the fitted distribution) is then provided to the expert to check the adequacy of the elicited distribution.

3.3.3 | Prior(s) for λ_e and γ_e

The final two parameters which we need to elicit distributions for are the two treatment parameters, λ_e and γ_e . We would not expect an expert to make judgements about these parameters directly. We instead follow usual practice of eliciting judgments about observable quantities,⁴² from which a prior for λ_e and γ_e can be inferred. Some possible choices of observable quantities are

- median survival time on the experimental treatment;
- survival probability at time t ; and
- greatest distance between survival curves and how big is this difference.

In practice, we have found experts have a preference to make judgements about hazard ratios, for example,

- hazard ratio at time t ; and
- maximum hazard ratio and when this occurs.

To elicit λ_e and γ_e , we require the expert to provide their beliefs for at least two of the above questions, which is likely to be a difficult task for the expert. We can simplify the elicitation task by making the assumption that $\gamma_e = \gamma_c$. We then have a piecewise-constant hazard ratio

$$\text{HR}(t) = \begin{cases} 1, & t \leq T \\ \left(\frac{\lambda_e}{\lambda_c}\right)^{\gamma_c}, & t > T \end{cases} \quad (11)$$

We can rearrange Equation (11) for the case when $t > T$ to obtain

$$\lambda_e = \lambda_c \text{HR}^{\frac{1}{\gamma_c}}. \quad (12)$$

Hence, conditional on λ_c and γ_c we can elicit a distribution for the hazard ratio for $t > T$, from which a distribution for λ_e can be derived. We make a standard modelling assumption that the treatment effect as described by the hazard ratio is independent of the control group response as determined by the parameters λ_c and γ_c . We investigate the implications of assumption $\gamma_e = \gamma_c$ in Section 5.

We denote the post-delay hazard ratio by HR^* (where it is assumed that S is true). We propose a prior

$$\text{HR}^* | S \sim \text{Gamma}(c, d). \quad (13)$$

Again, any non-negative distribution could be used for HR^* . As with T , we need questions that an expert would be willing to answer and from which we can identify values for c and d . We elicit judgements about the distribution HR^* by stating

“Suppose that the population survival curves separate. We now want you to consider your uncertainty about the hazard ratio once the experimental treatment begins to take effect.”

We would then use a standard method for eliciting a univariate distribution for HR^* , as described in Section 3.3.2.

Conditional on S and any data \mathbf{x}_{hist} related to control group survival, we assume a joint distribution of the form

$$\pi(T, \text{HR}^*, \lambda_c, \gamma_c | S, \mathbf{x}_{\text{hist}}) = \pi(\lambda_c, \gamma_c | S, \mathbf{x}_{\text{hist}}) \pi(T | S) \pi(\text{HR}^* | S), \quad (14)$$

with P_S completing the prior specification. For algorithmic convenience, if S is not true, we set $T = 0$ and $\text{HR}^* = \text{HR} = 1$. We do not expect S to be informative for the control group survivor function, so we assume $\pi(\lambda_c, \gamma_c | S, \mathbf{x}_{\text{hist}}) = \pi(\lambda_c, \gamma_c | \mathbf{x}_{\text{hist}})$. We have assumed an expert's judgements about $\text{HR}^* | S$ are conditionally independent of T given S . If the expert wanted

to incorporate dependence between these parameters, a more complicated elicitation method could be used, such as the SHELF extension method,⁴¹ illustrated in Holzhauer et al.¹⁷

The elicitation technique discussed above assumes a single expert, but it is likely that in practice multiple experts will be consulted. Eliciting a distribution from multiple experts typically involves either eliciting a distribution from each expert separately and then aggregating the results, or alternatively getting the experts to agree on a single distribution. The SHELF method involves a combination of the two: experts first make judgements independently, which are then shared with the group. Following a facilitated discussion, the experts are then asked to agree on a single distribution reflecting the perspective of a “Rational Impartial Observer”. Other methods for eliciting distributions from multiple experts are available.^{43,44}

3.4 | Computing assurance under the DTE model

We use these elicited distributions to calculate assurance for various sample sizes using Algorithm 1. This algorithm incorporates free parameters that can be adjusted to reflect operational constraints in a clinical trial: the control and treatment group sample sizes n_c and n_e , and the total number of required events E . We let E be a free parameter in the algorithm as it is common to run event driven survival trials. This is because, for a time-to-event endpoint, statistical information for the log-hazard ratio is a function of E and therefore attained power will be determined by the number of events observed at the analysis time. Changing these free parameters will have consequences for assurance, so it is important to consider different combinations of trial designs in order to find the one which best suits the needs of the sponsor. For example, for a fixed n_c and n_e , if we increase E (the number of events) we may increase assurance at the cost of needing to run the trial for longer.

The recruitment schedule and analysis technique in Algorithm 1 (and Algorithm 2 in Section 5.2) are left unspecified, as these choices are not part of the assurance methodology; they can be selected separately. In the example of Section 4, we use a Fleming-Harrington weighted log-rank test for the analysis (as there is high prior belief that the separation of the survival curves will be subject to a delay). By default, we assume uniform recruitment for 12 months.

To implement our methods we have developed an R Shiny app which is available both as an offline R package and hosted online. Instructions are provided in the Appendix. The Shiny app allows users to choose from two recruitment schedules: piecewise constant and power method (taken directly from the `nphRCT`⁴⁵ R package). Different testing approaches have been proposed for use in the non-proportional hazards setting (eg, max-combo test,¹¹ weighted log-rank tests,⁴⁶ difference in RMST,³⁷ and more⁴⁷). Our app offers two of these statistical tests: a standard log-rank test and a Fleming-Harrington weighted log-rank test taken from the `nph`³² R package). Finally, the elicitation process may inform refinements to the analysis plan.

4 | EXAMPLE

In this section, we illustrate the proposed method with a hypothetical example where we design a two-arm Phase III superiority trial to test whether a new drug is beneficial versus the current standard of care, docetaxel, in patients with advanced non-small-cell lung cancer (NSCLC). As the drug is in the IO area, we expect a DTE. The primary efficacy endpoint is OS, we assume uniform recruitment for 12 months, 1:1 allocation, and that the data will be analysed with a Fleming-Harrington weighted log-rank test with $\rho = 0$ and $\gamma = 1$, as we have a high probability that the treatment will be subject to a delay and we want to place more weight on late differences in the survival curves. We assume that the trial final analysis will take place when 80% of patients have died.

4.1 | Prior distribution(s) for the control parameters

There exists historical data on docetaxel, so we are able to use this to generate a prior distribution for control group parameters. In Bertsche et al³⁹ they found three trials in which docetaxel was used as the control in a clinical trial; ZODIAC,⁴⁸ REVEL⁴⁹ and INTEREST.⁵⁰ We also use the results from these three trials, but we use the published Kaplan-Meier curves

Algorithm 1. Calculating assurance when a DTE is likely to be present in a clinical trial

Inputs: sample sizes n_c and n_e , the elicited priors $\pi(\lambda_c, \gamma_c | \mathbf{x}_{\text{hist}})$, $\pi(T|S)$, $\pi(\text{HR}^*|S)$, the probability of the survival curves separating P_S , the number of events E (we require $E \leq n_c + n_e$) and the number of iterations N .

For $i = 1, \dots, N$:

1. sample $\lambda_{c,i}$ and $\gamma_{c,i}$ from $\pi(\lambda_c, \gamma_c)$;
2. set $\gamma_{e,i} = \gamma_{c,i}$;
3. sample survival times for the control group $x_{1,i}, \dots, x_{n_c,i}$ using $\lambda_{c,i}, \gamma_{c,i}$ (can use `rweibull()`);
4. sample u from `runif(0, 1)`
5. if $u < P_S$:
 - sample T_i and HR_i^* from $\pi(T)\pi(\text{HR}^*|S)$;
 - else:
 - set $T_i = 0$ and $\text{HR}_i^* = 1$;
6. transform HR^* to $\lambda_{e,i}$ using Equation 12;
7. sample survival times for the experimental treatment group $y_{1,i}, \dots, y_{n_e,i}$ using $T_i, \lambda_{e,i}$ and $\gamma_{e,i}$ (can use inversion sampling via Equation 8);
8. sample recruitment times $R_{1,i}, \dots, R_{n_c+n_e,i}$ from the pre-specified recruitment schedule;
9. add the survival times from each group to the recruitment times to obtain a pseudo event time $P_{1,i}, \dots, P_{n_c+n_e,i}$;
10. order the pseudo event times and define E_T to be the time at which E events have been observed;
11. remove any observation in which the recruitment time $R_{j,i} > E_T$;
12. censor any observation for which the pseudo event time $P_{j,i} > E_T$;
13. for any censored observation, redefine the survival time to be $E_T - R_{j,i}$;
14. perform the method of analysis on the data $x_{1,i}, \dots, x_{n_c,i}$ and $y_{1,i}, \dots, y_{n_e,i}$;
15. define $U_i = 1$ if the data give rise to a 'successful' outcome (0 otherwise).

The assurance is then estimated as

$$\hat{P}(R) = \frac{1}{N} \sum_{i=1}^N U_i.$$

to reconstruct the individual patient data.²⁵ The three Kaplan-Meier curves can be seen in Figure 3. Since survival in all three trials appears similar, we choose to pool the data from all three trials and use this to update non-informative priors for λ_c and γ_c , using Markov chain Monte Carlo (MCMC) to sample from the posterior distributions. The generated MCMC samples are then used as a prior distribution for the future trial of interest.

4.2 | Eliciting the prior distribution for the length of delay

We now need to elicit the expert's probability that the population survival curves separate at some point in time. Suppose the expert specifies this as 90%, that is $P_S = 0.9$. The expert is then asked for their uncertainty about the length of delay, given that the survival curves do separate. Suppose the expert's probability that the effect of the experimental treatment will be subject to a delay is 70%, that is, $P_{\text{DTE}} = 0.7$. The expert is then for about their beliefs about T , conditional on there being a delay. The expert provides a median of 4 months and two quartiles (25% and 75%) of 3 and 5 months, respectively. A $\text{Gamma}(a, b)$ distribution is fitted to these judgements, so that $D_{\text{delay}} = \text{Gamma}(7.29, 1.76)$. Combining these beliefs using Equation (10), we have the following mixture prior distribution

$$T = \begin{cases} 0, & \text{with probability } 0.3 \\ \text{Gamma}(7.29, 1.76), & \text{with probability } 0.7 \end{cases}$$

The fitted quartiles of a $\text{Gamma}(7.29, 1.76)$ distribution are 3.03, 3.95 and 5.05. These would be presented to the expert for feedback.

Algorithm 2. Calculating flexible assurance for when a DTE is likely to be present

Inputs: sample sizes n_c and n_e , the elicited priors $\pi(\lambda_c, \gamma_c | \mathbf{x}_{\text{hist}})$, $\pi(T|S)$, $\pi(\text{HR}^*|S)$, the probability of the survival curves separating P_S , the number of events E (we require $E \leq n_c + n_e$), the maximum trial length L_{max} , the number of initial samples M and the number of iterations N .

1. Initialise an empty matrix $A \in \mathbb{R}^{M \times t}$, where t is the length of a vector `time = seq(0, Lmax, by = 0.01)`;
2. for $j = 1, \dots, M$:
 - i sample $\lambda_{c,j}, \gamma_{c,j}$ from $\pi(\lambda_c, \gamma_c)$;
 - ii set $\gamma_{e,j} = \gamma_{c,j}$;
 - iii sample u from `runif(0, 1)`;
 - iv if $u < P_S$:
 - sample T_j, HR_j^* from $\pi(T, \text{HR}^*|S)$;
 - else:
 - set $T_j = 0, \text{HR}_j^* = 1$;
 - v transform HR_j^* to $\lambda_{e,j}$ using Equation 12;
 - vi use $T_j, \lambda_{e,j}, \gamma_{e,j}$, and the control parameters λ_c, γ_c to calculate the survival probability at each value of the time vector, using Equation 8;
 - vii fill the j 'th row of the matrix A with these survival probabilities.
3. For $i = 1, \dots, N$:
 - i sample $\lambda_{c,i}, \gamma_{c,i}$ from $\pi(\lambda_c, \gamma_c)$;
 - ii define F to be the time at which the survival probability in the control group equals 0.01 (using $\lambda_{c,i}, \gamma_{c,i}$ and Equation 5);
 - iii sample a survival probability, $s_{1,i}$, from the column in matrix A which corresponds to $0.25F$;
 - iv sample a survival probability, $s_{2,i}$ from the column in matrix A which corresponds to $0.6F$ (we require $s_{2,i} < s_{1,i}$);
 - v sample T_i from $\pi(T)$;
 - vi simultaneously solve these equations to find the best fitting values of $\lambda_{e,i}$ and $\gamma_{e,i}$ (can use `nleqslv()`);
 - vii sample survival times for the control group $x_{1,i}, \dots, x_{n_c,i}$ using the sampled $\lambda_{c,i}, \gamma_{c,i}$ (can use `rweibull()`);
 - viii sample survival times for the experimental treatment group $y_{1,i}, \dots, y_{n_e,i}$ using the sampled T_i , and $\lambda_{e,i}, \gamma_{e,i}$ (can use inversion sampling);
 - ix sample recruitment times $R_{1,i}, \dots, R_{n_c+n_e,i}$ from from the pre-specified recruitment schedule;
 - x add the survival times from each group to the recruitment times to obtain a pseudo event time $P_{1,i}, \dots, P_{n_c+n_e,i}$;
 - xi order the pseudo event times and define E_T to be the time at which E events have been observed;
 - xii remove any observation in which the recruitment time $R_{j,i} > E_T$;
 - xiii censor any observation for which the pseudo event time $P_{j,i} > E_T$;
 - xiv for any censored observation, redefine the survival time to be $E_T - R_{j,i}$;
 - xv perform the method of analysis on the data $x_{1,i}, \dots, x_{n_c,i}$ and $y_{1,i}, \dots, y_{n_e,i}$;
 - xvi define $U_i = 1$ if the data give rise to a 'successful' outcome (0 otherwise).

The assurance is then estimated as

$$\hat{P}(R) = \frac{1}{N} \sum_{i=1}^N U_i.$$

4.3 | Eliciting the prior distribution for the post-delay hazard ratio

The second quantity of interest is $\text{HR}^*|S$. Suppose the expert provides a median of 0.6 and two quartiles (25% and 75%) of 0.55 and 0.7, respectively. Again, we fit a $\text{Gamma}(c, d)$ to these judgements, so $D_{\text{HR}} = \text{Gamma}(29.6, 47.8)$. As per Equation (13), we have the following prior distribution

$$\text{HR}^*|S \sim \text{Gamma}(29.6, 47.8).$$

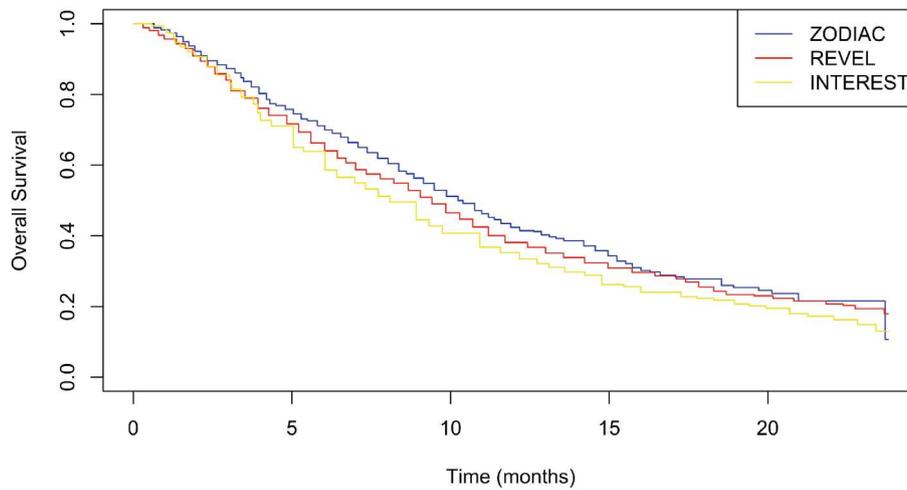


FIGURE 3 Reconstructed Kaplan-Meier curves for the docetaxel arm in three different trials: ZODIAC,⁴⁸ REVEL⁴⁹ and INTEREST.⁵⁰ We see that the three curves are similar and we assume exchangeability of the trials in our example.

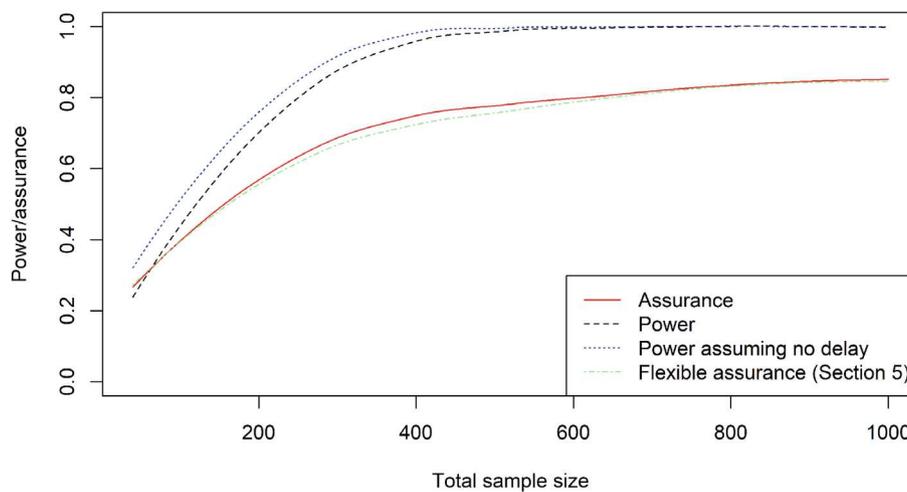


FIGURE 4 Power/assurance curves for the example given in Section 4. We see that the sample size required for 80% power/assurance is greatly different under the different scenarios and highlights the importance of including uncertainty in the design stage of a clinical trial.

The fitted quartiles of Gamma(29.6, 47.8) distribution are 0.54, 0.61 and 0.69, and again, this would be presented to the expert for feedback.

4.4 | Calculating assurance

We use these elicited prior distributions to calculate assurance for this example using Algorithm 1. In Figure 4, an assurance curve is plotted to inform sample sizes required for this clinical trial. Also seen in Figure 4 are three other power/assurance curves. The two power curves correspond to including no uncertainty in the parameters, with the control parameters, λ_c and γ_c , being the MLE from the three pooled data sets, as discussed in Section 4.1. The values for T and HR^* are the median values given by the experts. For one of the power curves, we have assumed that T is 0, and therefore does not account for the fact that the treatment effect may be subject to a delay. Also shown is an assurance curve, corresponding to a more flexible approach to calculating assurance, this approach is presented in Section 5.2. The distributions/values for the first three curves are found in Table 1. We kept the recruitment scheduling, analysis method, and so forth, as described at the start of Section 4, constant across all four scenarios.

TABLE 1 Distribution/values for the parameters, for the first three scenarios seen in Figure 4.

Calculation	λ_c	γ_c	T	HR*	P_S	P_{DTE}
Assurance	MCMC sample	MCMC sample	Ga(7.29, 1.76)	Ga(29.6, 47.8)	0.9	0.7
Power	0.074	1.21	4	0.6	1	1
Power assuming no delay	0.074	1.21	0	0.6	1	0

In Figure 4, we see that both of the power calculations are much more optimistic than the other two scenarios; we require far fewer patients for the same power, at all sample sizes. This highlights the importance of incorporating uncertainty into the trial parameters. However, we must reiterate, the assurance method is not simply used for setting sample sizes for the proposed trial. We anticipate the assurance method being used as one step in a thorough process to decide whether or not to go ahead with the trial, and if we do run the trial, define the characteristics of the proposed trial; length, number of patients, number of events, and so forth. For example, if we required a quicker trial, we may choose to decrease the number of events we need to observe before stopping the trial, E , but this will surely come at a cost of reducing the assurance/power seen. Therefore, it is important that a number of different trial designs are considered, and then assurance curves can be plotted to help inform the ultimate decision(s).

5 | SIMPLIFIED PRIOR DISTRIBUTION: DISCUSSION

To elicit the parameters in this scenario, we select a method that would be both effective and straightforward for the experts. As a result, we simplified the original parameterisation by fixing $\gamma_e = \gamma_c$. By doing so, we were able to focus on hazard ratios as the basis for our questioning, thus ensuring that the elicitation process was both easy and intuitive for the experts.

However, it's important to investigate the robustness of this simplification. The results of our investigation are presented in the following section. Finally, we provide an alternative method for calculating assurance. This alternative is designed to accommodate situations in which the aforementioned simplification may not be preferred.

5.1 | Robustness of the parameterisation

The investigation aimed to assess the impact of the simplification we introduced, $\gamma_e = \gamma_c$, into the model by comparing two parameterisation methods: Method A and Method B. Method A incorporated the simplification, while Method B allowed γ_e to vary. Historical data from three clinical trials (Checkmate 017,¹⁰ Checkmate 141,⁵¹ and Checkmate 017 and Checkmate 057 combined,⁵² the Kaplan-Meier plots for these trials are seen in Figure 5) with observed DTE were used

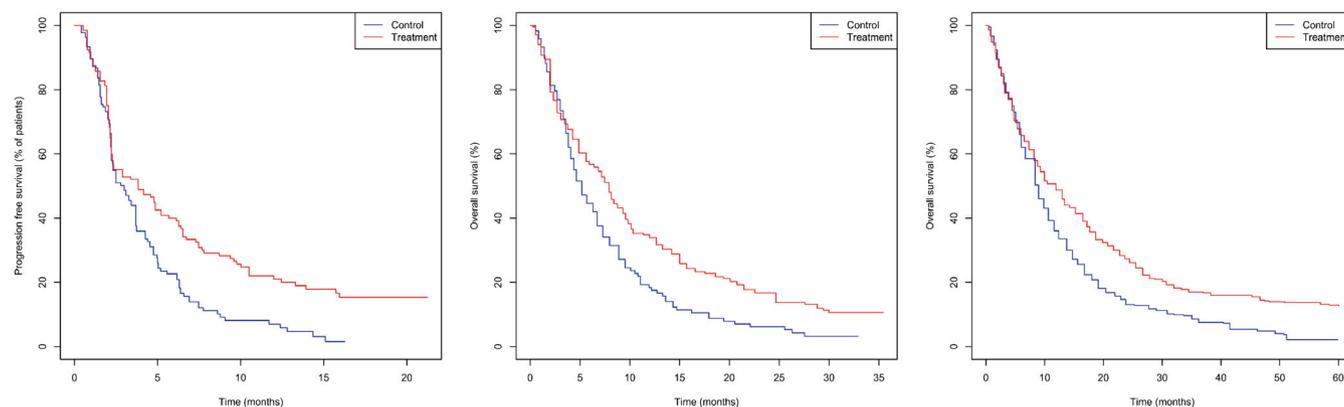


FIGURE 5 Kaplan-Meier plots for the data sets introduced in Section 5. For (A) the data set is from trial Checkmate 017,¹⁰ for (B) the data set is from trial Checkmate 141⁵¹ and for (C) the data set is from trials Checkmate 017 and Checkmate 057 combined.⁵²

TABLE 2 How each of the five parameters are estimated in both of the methods introduced in Section 5, Method A is the simplification we made in the elicitation process (MLE = maximum likelihood estimation).

Method	T	γ_c	λ_c	γ_e	λ_e
A	Visually	survreg(dist = "weibull")		γ_c	MLE
B				MLE	MLE

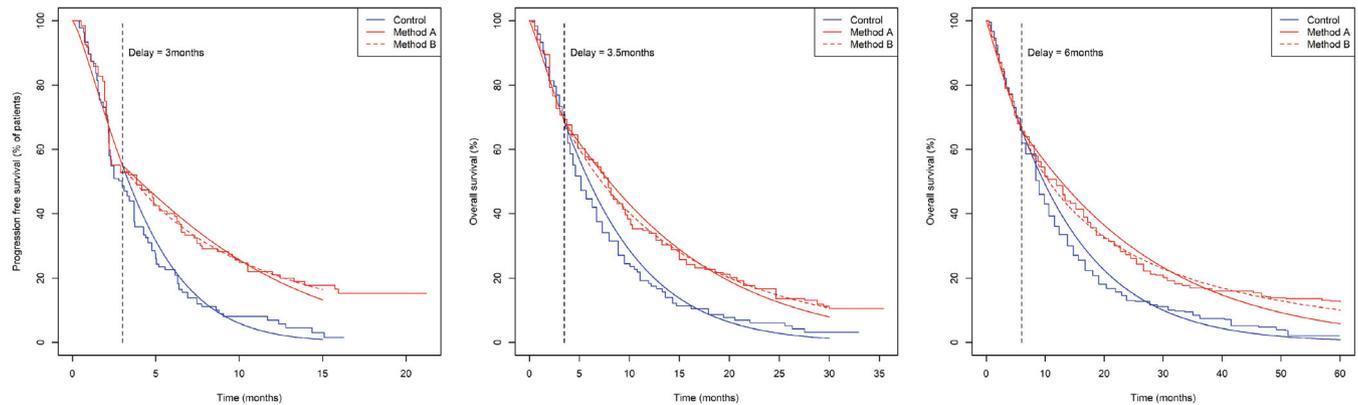


FIGURE 6 The same Kaplan-Meier plots as in Figure 5, with estimates for both Method A and B (as introduced in Section 5) overlaid. For each data set, T has been visually estimated. The control parameters, λ_c and γ_c , have been estimated from the data (the overlaid blue line). In Method A, the simplification has been made ($\gamma_e = \gamma_c$) and then λ_e has been estimated using a least squares procedure. In Method B, both the treatment parameters (λ_e, γ_e) have been simultaneously estimated.

to estimate the five unknown parameters (from Equations 5 and 8) using both methods. For clarity, Table 2 shows how the two methods estimate the five unknown parameters.

The estimated parametric survival curves generated by both methods are presented in Figure 6. Observing all three examples, we can see that the parametric treatment survival curve produced by Method B exhibits a marginally superior fit compared to the treatment curves derived from Method A. This is what we would intuitively expect, as Method B incorporates two free parameters, γ_e and γ_c , while Method A only employs a single free parameter, γ_c . However, despite Method B giving a better fit to the data, Method A still approximates the data well. These findings suggest that the simplification introduced by Method A would likely have minimal practical impact on real decision-making processes.

Power calculations were performed to quantify the impact of the difference in the fitted experimental treatment survival curves. The results, depicted in Figure 7, showed almost indistinguishable power curves for both methods across all three datasets. This suggests that, in these examples, the assumption $\gamma_e = \gamma_c$ did not lead to different practical outcomes.

5.2 | A more flexible approach to evaluating assurance

We have demonstrated that for the three historical trials considered, the simplification does not appear to have any practical implications. However, by making this simplification, the possible experimental treatment survival curves are constrained to align with the shape of the control survival curve. In Figure 8A, it can be observed that the experimental treatment survival curves seem to be ‘parallel’ to the control curve due to this simplification and the fixed shape parameters, $\gamma_e = \gamma_c$, being the same for both curves.

It is important to acknowledge that in certain trial designs, practitioners may feel uneasy about using this method, particularly if they believe that the experimental treatment survival curve would not align in a parallel manner to the control curve. In response to this concern, we have developed an alternative assurance calculation, referred to as Algorithm 2. Algorithm 2 aims to generate experimental treatment curves that would be obtained if we had not made the simplification and allowed the curves to be sampled from a Weibull(λ_e, γ_e) distribution after the delay occurred. The process is depicted in Figure 9.

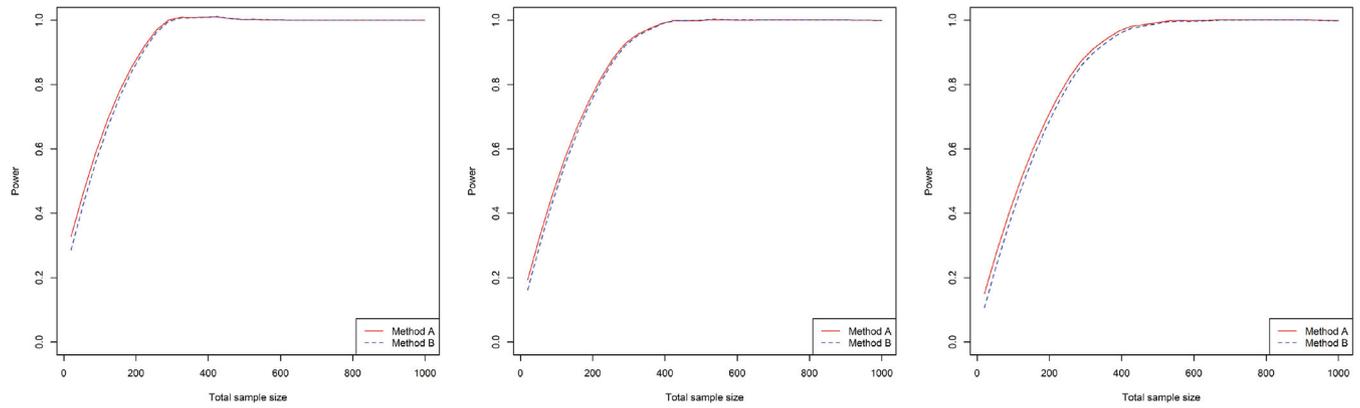


FIGURE 7 Power curves for the two methods considered in Section 5, for all three of the trials. We see that, in all three cases, the power curves are very similar to each other, thus indicating that the simplification does not seem make any practical difference in these examples. Method A is the simplification we made in the elicitation process.

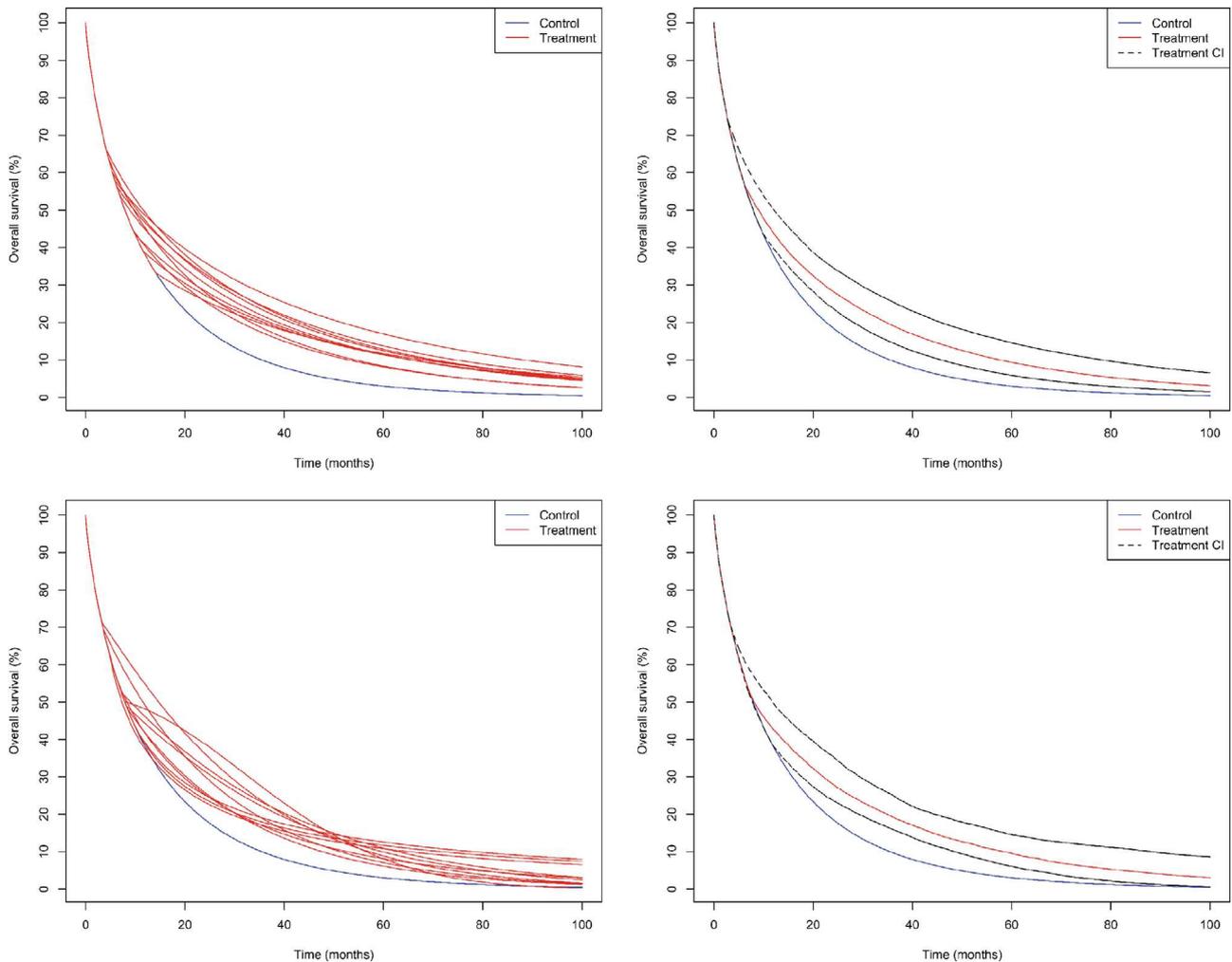


FIGURE 8 (A) Ten sampled experimental treatment survival curves, generated by Algorithm 1, (B) pointwise confidence intervals (0.1 and 0.9) for 500 of these sampled experimental treatment curves, (C) 10 sampled experimental treatment survival curves, generated by Algorithm 2, (D) pointwise confidence intervals (0.1 and 0.9) for 500 of these sampled experimental treatment curves.

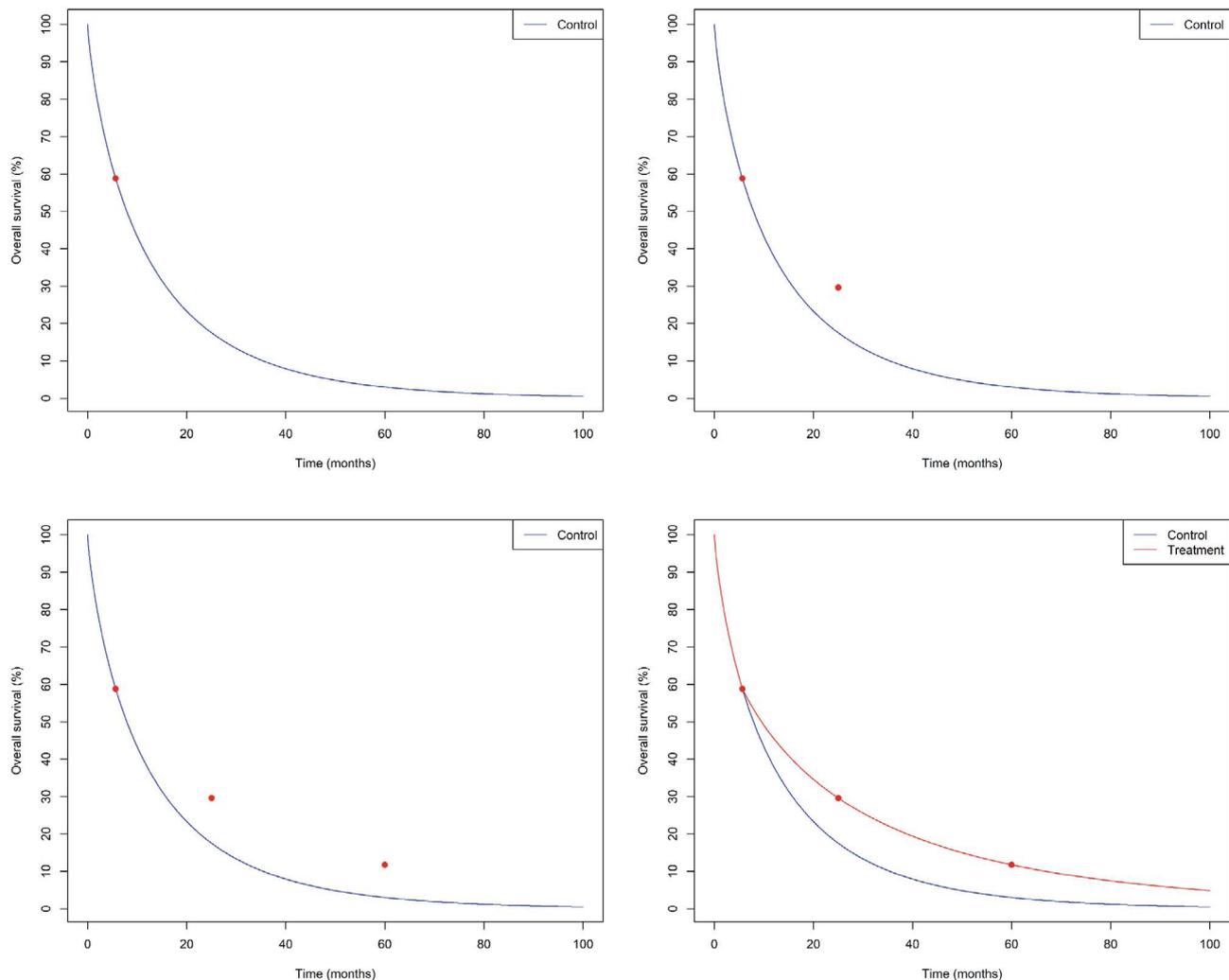


FIGURE 9 Showing the process of how an experimental treatment survival line is drawn for the flexible assurance. (A) A delay time is drawn from the prior distribution for T . (B) s_1 is sampled from all the survival probabilities at $0.25F$ through the trial, (C) s_2 is sampled from all the survival probabilities at $0.6F$ through the trial, (D) a piecewise Weibull distribution is fit to these sampled points.

To implement Algorithm 2, we utilize the elicited mixture prior distributions for T and $HR^*|S$. First, we sample a large number of experimental treatment curves, denoted as M . Next, we sample a value for T from the elicited prior distribution, as shown in Figure 9A. We define F as the time at which the control curve reaches a survival probability of 0.01. Then, we independently sample two survival probabilities, s_1 and s_2 , at $0.25F$ and $0.6F$ through the trial, respectively, as depicted in Figure 9B,C. The only condition imposed is that $s_1 > s_2$. Using T , s_1 and s_2 , we apply a least squares procedure to fit a Weibull distribution to these points, as illustrated in Figure 9D.

The sampled experimental treatment curves obtained from Algorithm 2 can be used to calculate assurance in the same manner as Algorithm 1. Figure 8C displays 10 sampled experimental treatment curves using this more flexible method. Additionally, Figure 8D shows the pointwise confidence intervals of the experimental treatment curves. Comparing these figures to Figure 8A,B, we observe that the pointwise confidence intervals are very similar, indicating that the sampled experimental treatment curves fall within the same boundaries. However, the alternative method allows for sampling a more diverse range of curves, providing increased flexibility.

We have implemented Algorithm 2 in the example presented in Section 4, and the results are seen in Figure 1. The flexible assurance curve closely resembles the assurance curve obtained using Algorithm 1. This demonstrates that the more flexible assurance method may not significantly impact decision-making, but it may make practitioners more comfortable if they believe that the imposed constraint is not representative of the potential experimental treatment curves

observed in practice. It is worth noting that the selection of time points ($0.25F$ and $0.6F$) for sampling survival probabilities in Algorithm 2 is somewhat arbitrary. These values were chosen to produce realistic experimental treatment survival curves. However, if more restrictive or flexible curves are desired, these two points could be adjusted accordingly (eg, closer together or further apart).

6 | SUMMARY

In conclusion, assurance calculations have emerged as a valuable tool in the design and analysis of clinical trials. By incorporating Bayesian principles and considering prior distributions for unknown parameters, assurance calculations provide a more realistic assessment of a trial's probability of success compared to traditional power calculations. This approach acknowledges the inherent uncertainties in clinical research and allows for the simulation of trial outcomes based on sampled prior distributions. Assurance calculations offer several advantages for trial design and decision-making. They assist in optimizing sample size, assessing risks, and evaluating the effectiveness of different trial setups, including the timing and number of planned interim analyses. Furthermore, assurance evaluations enable better-informed go/no-go decisions regarding study conduct, directing resources towards research programs with the highest expected impact for patients.

In the rapidly evolving field of immuno-oncology, assurance calculations have the potential to address challenges associated with time-varying or delayed treatment effects on time-to-event endpoints. We have extended the assurance method to include survival trials in which a delayed treatment effect is likely to occur. Overall, assurance calculations provide a robust framework for quantifying the probability of success in clinical trials while considering uncertainty. By incorporating Bayesian methods and accommodating complexities in trial design, assurance calculations contribute to more informed decision-making, improved trial design, and ultimately, more effective and impactful clinical research.

ACKNOWLEDGEMENTS

This work has been supported by a University of Sheffield EPSRC Doctoral Training Partnership (DTP) Case Conversion with Novartis Scholarship [project reference 2610753].

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article, as no new data were created or analysed in this study.

ORCID

James A. Salsbury  <https://orcid.org/0000-0002-2584-3640>

Steven A. Julious  <https://orcid.org/0000-0002-9917-7636>

REFERENCES

1. Spiegelhalter D, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med*. 1986;5(1):1-13. doi:10.1002/sim.4780050103
2. O'Hagan A, Stevens J, Campbell MJ. Assurance in clinical trial design. *Pharm Stat*. 2005;4(3):187-201. doi:10.1002/pst.175
3. Grieve AP. *Hybrid Frequentist/Bayesian Power and Bayesian Power in Planning Clinical Trials*. Boca Raton, FL: CRC Press LLC; 2022.
4. Hampson LV, Holzhauer B, Bornkamp B, et al. A new comprehensive approach to assess the probability of success of development programs before pivotal trials. *Clin Pharmacol Therapeut*. 2021;111(5):1050-1060. doi:10.1002/cpt.2488
5. Hampson LV, Bornkamp B, Holzhauer B, et al. Improving the assessment of the probability of success in late stage drug development. *Pharm Stat*. 2021;21(2):439-459. doi:10.1002/pst.2179
6. Götte H, Schüler A, Kirchner M, Kieser M. Sample size planning for phase II trials based on success probabilities for phase III. *Pharm Stat*. 2015;14(6):515-524. doi:10.1002/pst.1717
7. US Food and Drug Administration. *22 Case Studies where Phase 2 and Phase 3 Trials Had Divergent Results*. 2017. Accessed June 6, 2024. <https://www.fda.gov/media/102332/download>
8. Dallow N, Best N, Montague TH. Better decision making in drug development through adoption of formal prior elicitation. *Pharm Stat*. 2018;17(4):301-316. doi:10.1002/pst.1854
9. Crisp A, Miller S, Thompson D, Best N. Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. *Pharm Stat*. 2018;17(4):317-328. doi:10.1002/pst.1856
10. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *New Engl J Med*. 2015;373(2):123-135. doi:10.1056/nejmoa1504627

11. Mukhopadhyay P, Ye J, Anderson KM, et al. Log-rank test vs MaxCombo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in Immuno-oncology trials. *JAMA Oncol.* 2022;8(9):1294. doi:10.1001/jamaoncol.2022.2666
12. Rizvi NA, Cho BC, Reinmuth N, et al. Durvalumab with or without Tremelimumab vs standard chemotherapy in first-line treatment of metastatic non-small cell lung cancer. *JAMA Oncol.* 2020;6(5):661. doi:10.1001/jamaoncol.2020.0237
13. Herbst RS, Giaccone G, de Marinis F, et al. Atezolizumab for first-line treatment of PD-L1-selected patients with NSCLC. *New Engl J Med.* 2020;383(14):1328-1339. doi:10.1056/nejmoa1917346
14. Shitara K, Van Cutsem E, Bang YJ, et al. Efficacy and safety of Pembrolizumab or Pembrolizumab plus chemotherapy vs chemotherapy alone for patients with first-line, advanced gastric cancer: the KEYNOTE-062 phase 3 randomized clinical trial. *JAMA Oncol.* 2020;6(10):1571-1580. doi:10.1001/jamaoncol.2020.3370
15. International Council for Harmonisation. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R). 2019. Accessed June 6, 2024. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.eps
16. Jiménez JL, Stalbovskaya V, Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. *Pharm Stat.* 2018;18(3):287-303. doi:10.1002/pst.1923
17. Holzhauer B, Hampson LV, Gosling JP, et al. Eliciting judgements about dependent quantities of interest: the SHEffield ELicitation framework extension and copula methods illustrated using an asthma case study. *Pharm Stat.* 2022;21(5):1005-1021. doi:10.1002/pst.2212
18. Hampson LV, Whitehead J, Eleftheriou D, et al. Elicitation of expert prior opinion: application to the MYPAN trial in childhood Polyarteritis Nodosa. *PLoS One.* 2015;10(3):e0120981. doi:10.1371/journal.pone.0120981
19. Gasparini M, Di Scala L, Bretz F, Racine-Poon A. Some uses of predictive probability of success in clinical drug development. *Epidemiology biostatistics and public Health.* 2022;10:e8760-1-e8760-14. doi:10.2427/8760
20. Alhussain ZA, Oakley JE. Assurance for clinical trial design with normally distributed outcomes: eliciting uncertainty about variances. *Pharm Stat.* 2020;19(6):827-839. doi:10.1002/pst.2040
21. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health Care Evaluation.* Hoboken, NJ: Wiley; 2004.
22. Hiance A, Chevret S, Lévy V. A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial. *J Clin Epidemiol.* 2009;62(4):431-437.e2. doi:10.1016/j.jclinepi.2008.04.009
23. Ren S, Oakley JE. Assurance calculations for planning clinical trials with time-to-event outcomes. *Stat Med.* 2013;33(1):31-45. doi:10.1002/sim.5916
24. Azzolina D, Berchiolla P, Gregori D, Baldi I. Prior elicitation for use in clinical trial design and analysis: a literature review. *Int J Environ Res Public Health.* 2021;18(4):1833. doi:10.3390/ijerph18041833
25. Liu N, Zhou Y, Lee JJ. IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves. *BMC Med Res Methodol.* 2021;21(1):1-22. doi:10.1186/s12874-021-01308-8
26. Fine GD. Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Inf J.* 2007;41(4):535-539. doi:10.1177/009286150704100412
27. Xu Z, Zhen B, Park Y, Zhu B. Designing therapeutic cancer vaccine trials with delayed treatment effect. *Stat Med.* 2016;36(4):592-605. doi:10.1002/sim.7157
28. Zhang D, Quan H. Power and sample size calculation for log-rank test with a time lag in treatment effect. *Stat Med.* 2009;28(5):864-879. doi:10.1002/sim.3501
29. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials [published correction appears in *biometrics* 1988 Sep;44(3):923]. *Biometrics.* 1988;44(1):229-241.
30. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika.* 1990;77(4):853-864. doi:10.1093/biomet/77.4.853
31. Luo X, Turnbull BW, Cai H, Clark LC. Regression for censored survival data with lag effects. *Commun Stat Theory Methods.* 1994;23(12):3417-3438. doi:10.1080/03610929408831455
32. Ristl R, Ballarini NM, Götte H, Schüler A, Posch M, König F. Delayed treatment effects, treatment switching and heterogeneous patient populations: how to design and analyze RCTs in oncology. *Pharm Stat.* 2020;20(1):129-145. doi:10.1002/pst.2062
33. Sit T, Liu M, Shnaidman M, Ying Z. Design and analysis of clinical trials in the presence of delayed treatment effect. *Stat Med.* 2016;35(11):1774-1779. doi:10.1002/sim.6889
34. Mukhopadhyay P, Huang W, Metcalfe P, Öhrn F, Jenner M, Stone A. Statistical and practical considerations in designing of immuno-oncology trials. *J Biopharm Stat.* 2020;30(6):1130-1146. doi:10.1080/10543406.2020.1815035
35. Chen TT. Statistical issues and challenges in immuno-oncology. *Journal for ImmunoTherapy of Cancer.* 2013;1(1):18-26. doi:10.1186/2051-1426-1-18
36. Li W, Sophie Y-P, Chen RA. Estimation of delay time in survival data with delayed treatment effect. *J Biopharm Stat.* 2018;29(2):229-243. doi:10.1080/10543406.2018.1534857
37. Royston P, Parmar MK. 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(1):152-166. doi:10.1186/1471-2288-13-152
38. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics.* 2014;70(4):1023-1032. doi:10.1111/biom.12242
39. Bertsche A, Fleischer F, Beyersmann J, Nehmiz G. Bayesian phase II optimization for time-to-event data based on historical information. *Stat Methods Med Res.* 2017;28(4):1272-1289. doi:10.1177/0962280217747310
40. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain Judgements.* Hoboken, NJ: John Wiley & Sons; 2006.

41. Oakley JE, O'Hagan A. *SHELF: The Sheffield Elicitation Framework (Version 4)*. Sheffield, UK: School of Mathematics and Statistics, University of Sheffield; 2023. Accessed June 6, 2024. <https://shelf.sites.sheffield.ac.uk>
42. Kadane JB, Wolfson LJ. Experiences in elicitation [read before the Royal Statistical Society at a meeting on "elicitation" on Wednesday, April 16th, 1997, the president, professor a. F. M. Smith in the chair]. *Stat.* 1998;47(1):3-19. doi:10.1111/1467-9884.00113
43. Cooke R. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: Oxford Univ. Press; 2011.
44. Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC. A practical guide to structured expert elicitation using the IDEA protocol. Anderson B, ed. *Methods Ecol Evol.* 2017;9(1):169-180. doi:10.1111/2041-210x.12857
45. Magirr D, Barrott I. nphRCT: non-proportional hazards in randomized controlled trials. R package version 0.1.0. 2022. Accessed June 8, 2024. <https://CRAN.R-project.org/package=nphRCT>
46. Fleming TR, Harrington DP. A class of hypothesis tests for one and two sample censored survival data. *Commun Stat.* 1981;10(8):763-794. doi:10.1080/03610928108828073
47. Horiguchi M, Hassett MJ, Uno H. Empirical power comparison of statistical tests in contemporary phase III randomized controlled trials with time-to-event outcomes in oncology. *Clin Trials.* 2020;17(6):597-606. doi:10.1177/1740774520940256
48. Herbst RS, Sun Y, Eberhardt WE, et al. Vandetanib plus docetaxel versus docetaxel as second-line treatment for patients with advanced non-small-cell lung cancer (ZODIAC): a double-blind, randomised, phase 3 trial. *Lancet Oncol.* 2010;11(7):619-626. doi:10.1016/s1470-2045(10)70132-7
49. Garon EB, Ciuleanu TE, Arrieta O, et al. Ramucirumab plus docetaxel versus placebo plus docetaxel for second-line treatment of stage IV non-small-cell lung cancer after disease progression on platinum-based therapy (REVEL): a multicentre, double-blind, randomised phase 3 trial. *Lancet.* 2014;384(9944):665-673. doi:10.1016/s0140-6736(14)60845-x
50. Kim ES, Hirsh V, Mok T, et al. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. *Lancet.* 2008;372(9652):1809-1818. doi:10.1016/s0140-6736(08)61758-4
51. Yen C, Kiyota N, Hanai N, et al. Two-year follow-up of a randomized phase III clinical trial of nivolumab vs. the investigator's choice of therapy in the Asian population for recurrent or metastatic squamous cell carcinoma of the head and neck (CheckMate 141). *Head Neck.* 2020;42(10):2852-2862. doi:10.1002/hed.26331
52. Borghaei H, Gettinger S, Vokes EE, 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. *J Clin Oncol.* 2021;39(7):723-733. doi:10.1200/jco.20.01605
53. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2022. Accessed June 8, 2024. <https://www.R-project.org/>
54. Chang W, Cheng J, Allaire JJ, et al. Shiny: web application framework for R. R package version 1.7.2. 2022. Accessed June 8, 2024. <https://CRAN.R-project.org/package=shiny>

How to cite this article: Salsbury JA, Oakley JE, Julious SA, Hampson LV. Assurance methods for designing a clinical trial with a delayed treatment effect. *Statistics in Medicine.* 2024;1-18. doi: 10.1002/sim.10136

APPENDIX

An R⁵³ package, DTEAssurance, for implementing the methods described in this paper is available on GitHub, at <https://github.com/jamesalsbury/DTEAssurance>. The website also includes an illustration of using the package to replicate the examples in this paper. This package is installed with the commands

```
install.packages("devtools")
devtools::install_github("jamesalsbury/DTEAssurance").
```

An app for implementing these methods, produced with shiny,⁵⁴ can be used online at <https://jamesalsbury.shinyapps.io/DTEAssurance/>. A version of the app for offline use is included in the DTEAssurance package.