Assurance Methods for Adaptive Clinical Trials with a Delayed Treatment Effect

James Andrew Salsbury

School of Mathematical and Physical Sciences

The University of Sheffield

A thesis submitted in partial fulfilment for the degree of
Doctor of Philosophy

December 2025

Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: _____

Signature: _____

Date: _____

Acknowledgements

First and foremost, I would like to thank my university supervisors, Professor Jeremy Oakley and Professor Steven Julious. Your dedication, patience, and expert guidance have supported me from the very beginning, and I am deeply grateful for everything you have taught me.

I am also very grateful to my industry supervisor, Dr Lisa Hampson. Thank you for being such a welcoming host during my visits to Basel and for your invaluable insight throughout both my PhD and internship.

I would like to acknowledge the University of Sheffield and Novartis for providing the funding that made this research possible, including the opportunities to visit Basel and to collaborate with the wonderful team at Novartis.

Thank you to everyone who has made my time in Sheffield so memorable, both within the Hicks Building and beyond. In particular, thank you to the people in I8 for keeping me company over the past four years. Special thanks go to Dom, Eloise, Brad, and Jack. From the moment I arrived, you made me feel welcome, and my time in Sheffield has been immeasurably more enjoyable because of you all.

To my family, thank you for always believing in me and for giving me the foundations and confidence to pursue whatever I wanted to do. You have supported me throughout my education and beyond, and your encouragement has shaped both my academic path and my confidence in pursuing it.

Finally, thank you to Elen. You are the most incredible partner, supporting me every step of the way. Thank you for your patience, encouragement, and for moving to Sheffield so we could be together—and I promise I'll get a proper job now.

Abstract

Modern oncology clinical trials increasingly face challenges posed by delayed treatment effects (DTEs), where therapeutic benefits emerge only after an initial delay period. Conventional methods that assume proportional hazards often underestimate these effects, leading to underpowered or inefficient studies. This thesis develops a Bayesian framework that integrates assurance methods, expert elicitation, and adaptive design principles to improve the planning and evaluation of such trials.

The first part establishes assurance as a Bayesian alternative to traditional power calculations, incorporating parameter uncertainty through prior distributions. Structured expert elicitation is used to construct these priors, ensuring that clinical knowledge is captured transparently and quantitatively. The approach is then extended to survival models representing DTEs, allowing uncertainty in both delay duration and post-delay treatment effects to propagate through assurance calculations.

Building on this foundation, the thesis introduces adaptive design strategies, particularly group sequential and predictive approaches, that use elicited priors to inform interim decision rules. These methods enhance trial efficiency and ethical conduct while maintaining statistical validity. All methods are implemented in a freely available `R` package, `DTEAssurance`, together with two interactive Shiny applications that enable practitioners to design and evaluate complex trials in real time.

Finally, anonymised oncology datasets from the Vivli platform are analysed to characterise empirical patterns of DTEs in modern immunotherapy trials. Together, these developments provide a reproducible and practical framework for integrating expert knowledge and Bayesian reasoning into adaptive survival trial design, advancing the methodological foundations of clinical trial planning and bridging innovation with real-world application.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clinical trials are central to evidence generation in drug development, providing unbiased and regulatory-grade evaluations of safety and efficacy (Sackett et al., 1996; Friedman, 2015). The credibility of their conclusions depends critically on choosing an appropriate design and, in particular, ensuring that the planned sample size yields a meaningful probability of detecting a clinically relevant treatment effect. Traditionally, this probability is quantified through a frequentist power calculation, which requires single-point assumptions for quantities such as the effect size, nuisance parameters, and event rates (Jones et al., 2003). However, these quantities are rarely known with precision during trial planning, and reliance on fixed inputs can lead to underpowered or unnecessarily large studies.

Bayesian assurance provides an alternative approach by integrating the power function over prior distributions for the unknown parameters (O'Hagan, Stevens, et al., 2005). By accounting for uncertainty explicitly, assurance delivers a more realistic measure of a trial's chance of success and has gained traction in both industrial and regulatory contexts (Dallow et al., 2018). Despite its appeal, the application of assurance to time-to-event (TTE) trials introduces additional complexities, since both the underlying hazard functions and their temporal behaviour critically influence information accrual and statistical power.

TTE trials follow patients until the occurrence of an event or censoring, and are commonly analysed using survival models and log-rank based methods (Machin et al., 2006; Kalbfleisch and Prentice, 2002). A pervasive assumption in their design is that of proportional hazards (PH), under which the hazard ratio comparing treatment and control is constant over time. In many therapeutic areas, especially immuno-oncology (IO), this assumption is frequently violated due to treatment effects that are delayed, diminish, or vary over time (Mukhopadhyay et al., 2022). Non-proportional hazards (NPH) can materially distort the relationship between sample size, information, and

power, particularly in group sequential and adaptive monitoring frameworks (Fleming and Harrington, 1981; Ristl et al., 2021).

One common form of NPH is the *delayed treatment effect* (DTE), in which the hazards for the two arms are similar for an initial period before diverging. Delayed effects have been documented in numerous IO trials, often driven by biological mechanisms such as immune activation and tumour microenvironment modulation (Chen, 2013). Designing trials where a DTE is plausible requires specifying: (i) whether a delay exists; (ii) the likely duration of the delay; and (iii) the magnitude of the eventual treatment effect. Misspecifying any of these can substantially bias power calculations, influence the timing of interim analyses, and compromise the reliability of operating characteristics (Fine, 2007; Chen, 2013). Yet in practice, early-phase data seldom provide adequate precision on these quantities.

This motivates the use of *expert elicitation* to capture relevant clinical and biological knowledge. Elicitation methodologies offer structured procedures for translating expert beliefs into statistically coherent prior distributions. While elicitation has been applied to treatment effects and event rates in other contexts (Dias et al., 2018), there is currently no established framework for eliciting both delay-related parameters and post-delay hazard ratios in survival trials. Such priors are essential for obtaining realistic assurance calculations when DTEs are anticipated, yet existing methods do not address this gap.

The design of adaptive clinical trials introduces further complexity. Adaptive designs allow pre-specified modifications, such as early stopping or sample size re-estimation, based on accumulating data (Pallmann et al., 2018). Statistical validity requires careful characterisation of information accrual, interim monitoring boundaries, and Type I error control (Jennison and Turnbull, 2000; U.S. Food and Drug Administration, 2019). Delays in treatment effect emergence can disrupt these operating characteristics, misalign information times, and affect predictive probabilities used for interim decision-making. Despite a growing literature on adaptive designs with survival endpoints, no existing framework integrates elicited delay-related priors into predictive or assurance-based evaluation of adaptive strategies.

1.1 Research Question and Objectives

This thesis addresses these methodological gaps by developing an approach that combines expert elicitation, Bayesian assurance, and adaptive design principles for survival trials with potential delayed treatment effects. The central research question guiding this work is:

> *"How can adaptive time-to-event trials be designed to incorporate elicited prior information about delayed treatment effects and to quantify probability of success under these assumptions?"*

To answer this question, the thesis pursues five objectives:

1. **To develop an elicitation framework tailored to delayed treatment effects.** This includes eliciting expert beliefs about the delay time and post-delay hazard ratio, and translating these beliefs into coherent prior distributions.

2. **To extend assurance methodology to incorporate these elicited priors.** This enables realistic probability-of-success calculations for fixed designs where delayed effects are plausible.

3. **To adapt these methods for use in adaptive trial designs.** This involves computing assurance and predictive probabilities for adaptive strategies, supporting robust interim decision-making.

4. **To evaluate the practical implications of delayed treatment effects using real clinical trial data.** This involves analysing completed oncology trials to assess empirical evidence of delay, compare delayed-effect models with proportional hazards alternatives, examine the impact of delay on interim monitoring decisions, and quantify the loss of statistical power when delayed effects are ignored at the design stage.

5. **To implement the methodology in a publicly available R package.** The software provides practitioners with tools for elicitation, prior construction, assurance computation, simulation, and evaluation of adaptive designs under delayed effects.

Together, these contributions advance a coherent methodological framework for planning both fixed and adaptive survival trials in settings where delayed treatment effects are expected.

1.2 Thesis Roadmap

The remainder of this thesis is organised to develop, justify, and apply a framework for designing survival trials with delayed treatment effects using elicited priors, Bayesian assurance, and adaptive decision rules.

Chapter 2 introduces the theoretical foundations of Bayesian assurance and its role in clinical trial design. The chapter reviews existing approaches for integrating parameter uncertainty into probability of success calculations and highlights methodological limitations when applied to time-to-event endpoints. This establishes the motivation for an extension of assurance, looking at anticipated delayed treatment effects.

Chapter 3 describes expert elicitation techniques, which are central to specifying prior distributions in assurance calculations. The chapter outlines structured methodologies for capturing expert beliefs and demonstrates how these can be formally integrated into Bayesian trial planning.

Chapter 4 extends the assurance framework to settings with delayed treatment effects (DTEs). It discusses alternative parameterisations for representing delay-time mechanisms, proposes a framework for eliciting corresponding prior distributions, and illustrates how these priors can be used to compute assurance for trials without interim analyses.

Chapter 5 reviews the principles of adaptive clinical trial design, highlighting the statistical and operational advantages of adaptivity as well as the associated methodological challenges. The chapter establishes the foundations for incorporating adaptive elements into assurance-based design.

Chapter 6 develops methods for incorporating elicited delay-related priors into adaptive design evaluation. The chapter presents predictive probabilities, adaptive monitoring rules, and assurance calculations for group sequential and other adaptive structures. Simulation studies illustrate how prior uncertainty in delay and treatment effect informs interim decisions and expected operating characteristics.

Chapter 7 provides an empirical assessment of delayed treatment effects using seven oncology trials obtained via the *Vivli* data-sharing platform. Individual participant data are used to quantify delay patterns, evaluate survival models, and assess the performance of adaptive monitoring procedures under real-world conditions. These analyses demonstrate the practical relevance of the proposed methodology and highlight the consequences of ignoring delayed effects at the design stage.

Finally, Chapter 8 synthesises the main findings, revisits the research objectives, and summarises the methodological contributions of the thesis. It concludes with recommendations for the design and evaluation of survival trials where delayed treatment effects are plausible and identifies directions for future methodological research.

In summary, the thesis develops an integrated set of tools, elicitation procedures, assurance methods, adaptive design extensions, and accompanying software, to support transparent, data-informed, and biologically grounded planning of clinical trials with delayed treatment effects.

1.3 Publications and Software

The research presented in this thesis has resulted in the following publications:

1. Salsbury JA, Oakley JE, Julious SA, Hampson LV. *Assurance methods for designing a clinical trial with a delayed treatment effect. Statistics in Medicine.* 2024; 43(19): 3595–3612. doi: 10.1002/sim.10136.
2. Salsbury JA, Oakley JE, Julious SA, Hampson LV. *Adaptive clinical trial design with delayed treatment effects using elicited prior distributions. arXiv preprint* 2025; arXiv:2509.07602. Available from: https://arxiv.org/abs/2509.07602. [Under revision at *Pharmaceutical Statistics*.]

To facilitate implementation and reproducibility of the methodologies developed in this thesis, an open-source `R` (R Core Team, 2024) package has been developed and released on CRAN (Salsbury, 2025):

- Salsbury J (2025). `DTEAssurance`: Assurance Methods for Clinical Trials with a Delayed Treatment Effect.
  `R` package version 1.0.1. Available from https://CRAN.R-project.org/package=DTEAssurance.

The `DTEAssurance` package provides functions for computing Bayesian assurance in settings where delayed treatment effects may arise, including both fixed and adaptive trial designs. It includes tools for prior elicitation and simulation-based evaluation of design operating characteristics.

In addition, two interactive Shiny (Chang et al., 2025) applications are included to support practical use and exploration of the proposed methods:

1. `DTEAssurance::assurance_shiny_app()`: an interactive interface for exploring the assurance framework described in Chapter 4, allowing users to specify prior distributions, delay models, and clinical trial parameters, and to visualise the corresponding operating characteristics associated to such a design: assurance, average sample size and average duration.
2. `DTEAssurance::assurance_adaptive_shiny_app()`: an interactive tool implementing the adaptive group sequential methodologies presented in Chapter 6, enabling users to assess adaptive decision criteria under information timings and stopping rules, under delayed effects.

Together, these outputs provide an integrated framework for the design, evaluation, and communication of clinical trials that accounts for delayed treatment effects using assurance methods.

Chapter 2

Assurance Methods

2.1 Introduction

This chapter introduces the concept of *assurance* in clinical trial design, a Bayesian analogue to the traditional notion of statistical power. Classical power calculations require the specification of fixed values for unknown parameters such as the treatment effect. In practice, especially during early planning, these quantities are uncertain, and power calculations based on single-point assumptions can give an overly optimistic or pessimistic view of a trial's operating characteristics. Assurance addresses this limitation by incorporating prior uncertainty about the treatment effect to provide an *unconditional* probability that the planned trial will achieve its statistical objective.

The chapter begins by reviewing conventional frequentist power calculations and highlighting their sensitivity to misspecification of design parameters. We then introduce assurance formally, showing how it extends the classical framework by averaging the power function over a prior distribution for the treatment effect and other relevant quantities. Illustrative examples are used to demonstrate how assurance and power can diverge when uncertainty is substantial. Finally, we discuss practical considerations for constructing prior distributions, laying the groundwork for the elicitation methods developed in the following chapter and for the integration of assurance within adaptive trial design later in the thesis.

2.2 Aims of the Chapter

The purpose of this chapter is to establish the theoretical and practical foundation for using assurance methods in clinical trial design. Assurance offers a Bayesian framework for quantifying the probability of trial success while accounting for uncertainty in treatment effects. The specific aims of the chapter are to:

1. Review classical power calculations for fixed trial designs and highlight their limitations.
2. Define and derive assurance for fixed trial designs as a Bayesian generalisation of power.
3. Compare assurance and traditional power to illustrate their conceptual and practical differences.
4. Demonstrate the application of assurance through analytical and simulation-based examples.
5. Discuss approaches to constructing prior distributions that underpin assurance calculations.

2.3 Background

The primary objective of a two-arm superiority clinical trial—comparing an investigational treatment, which refers to the new intervention being evaluated, to a control group, which receives either a placebo or the current standard of care—is to test the null hypothesis of no treatment effect,

$$H_0 : \delta = 0,$$

where δ denotes the true treatment effect. Depending on the endpoint, δ may represent a difference in mean response, a log-hazard ratio, or another clinically interpretable effect measure. In survival trials, for example, δ is typically the log-hazard ratio comparing the investigational treatment to control.

The alternative hypothesis is

$$H_1 : \delta > 0,$$

reflecting the expectation that the experimental treatment provides benefit relative to control. Hypothesis testing proceeds by evaluating the extremity of the observed test

statistic under the sampling distribution implied by H_0; if this probability (the p-value) falls below a pre-specified significance level, H_0 is rejected in favour of H_1.

Although many confirmatory trials employ two-sided tests to symmetrically control Type I error, clinical and regulatory interest in superiority settings typically concerns only the possibility that the investigational treatment is more effective than control. A result favouring control does not constitute evidence of efficacy. For this reason, this thesis adopts one-sided hypothesis tests, typically conducted at the 2.5% significance level.

Let x denote the data collected from patients randomised between two treatment arms. At the conclusion of the trial, a test statistic $T(x)$ is computed and compared with a pre-specified critical region C, defined such that the Type I error rate is controlled at level α:

$$P(T(x) \in C \mid H_0) = \alpha.$$

If $T(x)$ falls within C, the null hypothesis H_0 is rejected at the $100\alpha\%$ significance level.

When determining the required sample size, the alternative hypothesis is typically specified as

$$H_1 : \delta = \delta_0,$$

where $\delta_0 > 0$ denotes the clinically meaningful effect size. Under this assumption, there exists a probability β of committing a Type II error—i.e., failing to reject H_0 when δ_0 is true. The power of the study, $1 - \beta$, represents the probability of detecting a statistically significant effect if δ_0 is indeed the true value. The sample size is therefore chosen to achieve the desired power at the specified δ_0 and Type I error rate α:

$$P(T(x) \in C \mid \delta = \delta_0) = 1 - \beta. \tag{2.1}$$

However, as several authors have noted (Spiegelhalter, Freedman, and Blackburn, 1986; Grieve et al., 1991; Parmar et al., 1994; Spiegelhalter, Freedman, and Parmar, 1994; Berry, 1993), the power calculation in Equation (2.1) is conditional on a single assumed treatment effect δ_0. In practice, δ is unknown at the design stage and subject to considerable uncertainty. If a prior distribution $\pi(\delta)$ is available to represent current beliefs about δ, a more realistic measure of the probability of trial success can be obtained by averaging the conditional power over this distribution:

$$\text{Assurance} = \int_{\delta} P(T(y) \in C \mid \delta)\, \pi(\delta)\, \mathrm{d}\delta. \tag{2.2}$$

This quantity, sometimes referred to as the *average power* (Spiegelhalter and Freedman, 1986), provides an unconditional probability of achieving statistical significance under the specified design.

The terminology surrounding Equation (2.2) has varied across the literature. It has been described as the *average success probability* (Chuang-Stein, 2006), *assurance* (O'Hagan, Stevens, et al., 2005), *expected power* (Gillett, 1994), *Bayesian predictive power* (Harari et al., 2021; Rufibach et al., 2016), and *(predictive) probability of success (PoS)* (Götte et al., 2020; He et al., 2012; Hampson, Bornkamp, et al., 2022; Grieve, 2022; Gasparini et al., 2013; Saint-Hilary et al., 2019).

Although the term *probability of success* is sometimes used interchangeably with assurance, its interpretation in industry is often broader, encompassing the overall likelihood of success across an entire development programme (Hampson, Holzhauer, et al., 2022; Hampson, Bornkamp, et al., 2022). To maintain clarity, this thesis adopts the term *assurance* to refer specifically to the probability that a single planned clinical trial will achieve statistical significance under its prespecified design.

2.4 Normal–Normal Example

To illustrate the analytical formulation of assurance in a tractable setting, we begin with a case in which both the outcome and the prior distribution of the treatment effect are normally distributed. This conjugate Normal–Normal model provides a simple but insightful framework for understanding how assurance extends the classical power calculation by incorporating prior uncertainty about the treatment effect.

Assume a two-arm study design in which n patients are randomised to each arm (treatment and control), and the primary endpoint is continuous. Let $\hat{\delta}$ denote the observed difference in sample means between the treatment and control groups. Under the assumptions of independent observations, a known and common population variance σ^2, and equal allocation to arms, the sampling distribution of $\hat{\delta}$ is normal with mean δ (the true treatment effect) and variance $2\sigma^2/n$:

$$\hat{\delta} \sim \mathcal{N}\left(\delta, \frac{2\sigma^2}{n}\right). \tag{2.3}$$

We consider the one-sided hypothesis test

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta > 0.$$

In practice, the power function is evaluated at a specific clinically relevant effect size, denoted $\delta_0 > 0$. This corresponds to computing

$$\text{Power}(n) = \Phi\left(Z_{1-\alpha} + \frac{\sqrt{n}\,\delta_0}{\sqrt{2}\sigma}\right), \tag{2.4}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution, and Z_p is the p-quantile of the standard normal distribution (Cohen, 1988). Equation (2.4) gives the usual one-sided power for comparing two means with known variance.

Solving Equation (2.4) for the required per-arm sample size n to detect a specified treatment effect δ_0 with power $1-\beta$ yields (Julious, 2009):

$$n = \frac{2\sigma^2 \left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\delta_0^2}. \tag{2.5}$$

This expression forms the basis for conventional frequentist sample size determination.

To evaluate the *assurance*, the unconditional probability of trial success accounting for prior uncertainty in the treatment effect, we now consider the predictive distribution of $\hat{\delta}$, obtained by marginalising over a prior distribution for δ.

Following the approach of Berry, 1996, we assume a conjugate Normal prior on the treatment effect:

$$\delta \sim \mathcal{N}\left(\delta_0, \frac{2\sigma^2}{n_0}\right),$$

where δ_0 is the prior mean, here taken to coincide with the clinically relevant effect size used in the power calculation. In practice, δ_0 may reflect a point estimate from prior evidence, the MCID, or a target effect size, and the choice should be made explicit and justified. The parameter n_0 represents the *effective sample size per arm* of the prior. This parameterisation allows the prior to be interpreted as if it were based on n_0 hypothetical observations (Spiegelhalter, Abrams, et al., 2004).

Under this prior, the predictive distribution of the observed treatment effect is analytically tractable:

$$\hat{\delta} \sim \mathcal{N}\left(\delta_0, 2\sigma^2\left(\frac{1}{n_0} + \frac{1}{n}\right)\right). \tag{2.6}$$

Equation (2.6) captures both prior uncertainty (through n_0) and future sampling variability (through n).

For a one-sided hypothesis test at significance level α, the null hypothesis $H_0 : \delta = 0$ is rejected if the observed effect exceeds the critical value:

$$\hat{\delta} > \frac{\sqrt{2}\sigma}{\sqrt{n}} Z_{1-\alpha}. \tag{2.7}$$

Combining Equations (2.6) and (2.7), the *assurance*, the probability that a future observed treatment effect exceeds the critical value for rejection, is given by:

$$\text{Assurance}(n) = \Phi\left(\sqrt{\frac{n_0}{n_0+n}}\left(\frac{\delta_0\sqrt{n}}{\sqrt{2}\sigma} + Z_{1-\alpha}\right)\right). \tag{2.8}$$

This expression defines assurance as a function of the planned sample size n and the effective prior information n_0.

An important property of Equation (2.8) is its *asymptotic upper bound*. As the sample size increases ($n \to \infty$), the assurance converges to:

$$\lim_{n\to\infty} \text{Assurance}(n) = \Phi\left(\frac{\sqrt{n_0}\,\delta_0}{\sqrt{2}\sigma}\right), \tag{2.9}$$

which corresponds to the prior probability that the treatment effect exceeds the decision threshold.

This upper bound highlights a key distinction between assurance and power. Increasing the sample size n cannot raise the assurance beyond this limit, because assurance accounts for total predictive uncertainty—comprising both future sampling variability and prior uncertainty. While the sampling variance decreases with increasing n, the prior variance remains constant, imposing a fixed upper bound. In contrast, power is conditional on a fixed effect size and approaches 1 as n becomes large.

2.4.1 Example

To illustrate, consider a Phase III clinical trial to evaluate a new antihypertensive drug. The study is designed as a two-arm parallel trial with equal allocation. Based on prior clinical evidence, the anticipated treatment effect is $\delta_0 = 5$ and the known standard deviation is $\sigma = 10$.

Using Equation (2.5), with a one-sided Type I error rate $\alpha = 0.025$ and Type II error rate $\beta = 0.2$ (80% power), the required sample size is 62.79 patients per group (rounded to 63).

To move beyond the fixed-effect assumption, we incorporate prior uncertainty about the treatment effect. Suppose the prior corresponds to an effective sample size of $n_0 = 4$ patients per arm, implying a prior distribution of $\mathcal{N}(5, 50)$. Substituting into Equation (2.8), we obtain an assurance of approximately 58.2%, notably lower than the nominal power of 80%. The asymptotic upper bound from Equation (2.9) is approximately 76.0%. Figure 2.1 shows power and assurance curves for different sample sizes.

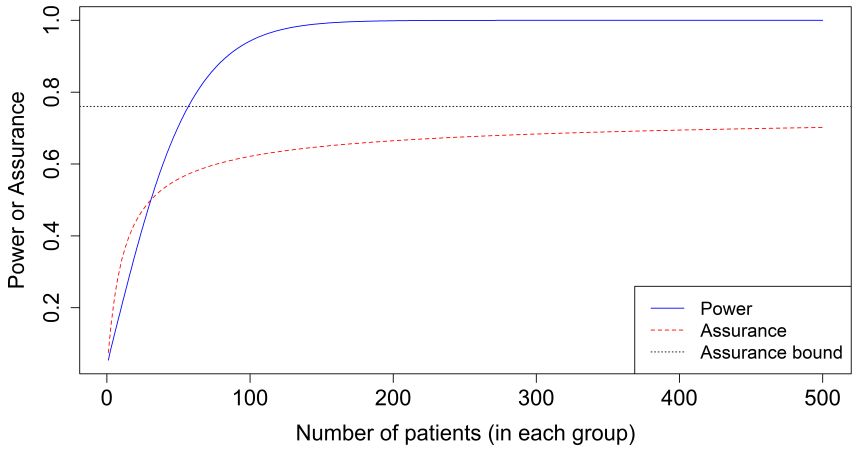


Figure 2.1: *Power and assurance curves for the example given in Section 2.4. The dashed line indicates the upper bound on assurance, showing that even large increases in sample size yield diminishing returns once prior uncertainty dominates.*

To better interpret the achievable gain in assurance, we can consider the *Normalised Assurance* (Muirhead and Soaita, 2012) or *scaled assurance* (Alhussain and Oakley, 2020), which rescales the assurance relative to its theoretical maximum:

$$\text{Normalised Assurance} = \frac{\Phi\left(\sqrt{\frac{n_0}{n_0+n}}\left(Z_{1-\alpha} + \frac{\sqrt{n}\,\delta_0}{\sqrt{2}\sigma}\right)\right)}{\Phi\left(\frac{\sqrt{n_0}\,\delta_0}{\sqrt{2}\sigma}\right)}. \tag{2.10}$$

For 63 patients per group, the normalised assurance is approximately 76.5%, providing a useful measure of the marginal benefit of additional sample size.

Throughout this work, comparisons with power are made using assurance rather than normalised assurance. Since assurance and power both lie on the same probability scale, this permits direct and intuitive comparison between the two metrics. Normalised assurance is a useful diagnostic for understanding the marginal benefit of additional sample size relative to the theoretical maximum, but rescaling removes this comparability. Furthermore, while the asymptotic upper bound on assurance is available analytically in the Normal–Normal setting, in more complex trial designs — such as those involving non-normal outcomes, cluster randomisation, or adaptive structures — the maximum assurance has no closed form and would itself require estimation via simulation, making normalised assurance less straightforward to compute in practice.

A further comparison between power and assurance can be made using their predictive distributions. Figure 2.2 shows the predictive distributions for power (Equation

(2.3)) and assurance (Equation (2.6) across a range of sample sizes, alongside the corresponding critical value (Equation (2.7)). Key observations include:

1. The area under the curve to the right of the critical value corresponds to the value of power or assurance.

2. As sample size increases, the variance of both distributions decreases. For power, this variance tends to zero as $n \to \infty$. For assurance, it converges to $2\sigma^2/n_0$, explaining the upper bound.

3. Assurance may exceed power at small sample sizes because of the incorporation of prior uncertainty and the resulting differences in critical thresholds.

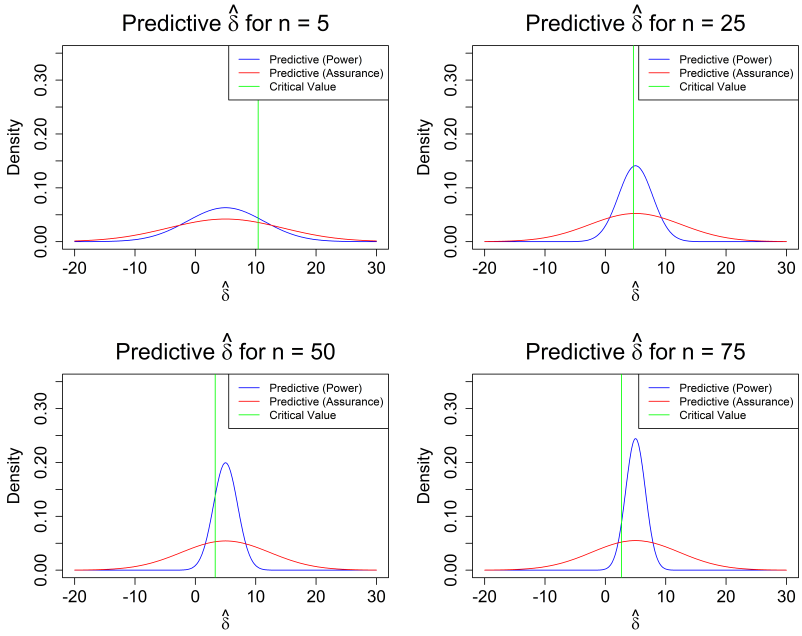


Figure 2.2: *Predictive distributions for power and assurance across varying sample sizes, demonstrating how the assurance distribution retains non-zero variance as n grows due to residual prior uncertainty.*

This Normal–Normal example clearly demonstrates how assurance generalises classical power by incorporating prior uncertainty and reveals its characteristic upper bound. In more realistic trial settings, such as those involving non-normal outcomes or complex test statistics, analytical expressions like Equation (2.8) are not available. The following section therefore introduces simulation-based methods for estimating assurance in such scenarios.

2.5 Simulation-Based Assurance

In Section 2.4, analytical expressions for assurance were obtainable because the Normal Normal model yields closed-form solutions under conjugacy. In realistic clinical trial settings, however, such tractability is exceptional. Designs involving non-normal endpoints, unknown variances, interim adaptations, or complex data structures typically preclude analytical derivation of either the power function or its prior weighted analogue. Even modest deviations from the idealised setting, such as non-linear test statistics, covariate adjustment, or joint uncertainty in multiple nuisance parameters, render closed-form assurance calculations infeasible.

In these situations, Monte Carlo simulation provides a flexible and widely applicable framework for estimating power and assurance. By repeatedly generating data under specific parameter values and propagating uncertainty through prior distributions, simulation methods allow the evaluation of operating characteristics for trials that lie far outside the scope of analytical theory. They are therefore indispensable for practical trial design, particularly in the presence of non-standard estimators, time-to-event endpoints, missing data mechanisms, or adaptive decision rules (Burton et al., 2006; Arnold et al., 2011; Morris et al., 2019). Simulation-based assurance thus forms the backbone of the methodology developed in subsequent chapters, enabling probability-of-success evaluation for models and designs that do not admit closed-form solutions.

Algorithm 1 outlines the standard Monte Carlo procedure for computing *power*. A data-generating model is first specified under the assumed true effect size. Repeated sampling of datasets and evaluation of the test statistic across N iterations yields an empirical estimate of the probability of rejecting the null hypothesis. As N increases, the Monte Carlo error diminishes, providing an accurate estimate of the study's operating characteristics.

Algorithm 1 Simulation procedure for estimating power. Repeated sampling of data under a fixed effect size yields the frequentist probability of rejecting H_0 at a specified significance level.

1: **Inputs:** Number of iterations N; control and treatment data generation models with assumed parameter values under H_1; critical value c_α under H_0
2: **for** $i = 1$ **to** N **do**
3: Simulate control and treatment group data under H_1
4: Compute test statistic t_i
5: Set $U_i = \mathbb{1}(t_i > c_\alpha)$ ▷ 1 if H_0 rejected, 0 otherwise
6: **end for**
7: Estimate power, where R denotes the event of rejecting H_0:

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

Algorithm 2 presents the corresponding simulation framework for *assurance*. The key distinction from power simulation lies in the first step: before generating data, parameter values are drawn from their prior distributions. This sampling step introduces epistemic uncertainty, ensuring that the simulation reflects the full predictive distribution of future trial outcomes. The result is an unconditional estimate of the probability of trial success, integrating over both parameter and data uncertainty.

Algorithm 2 Simulation procedure for estimating assurance. Sampling parameters from their prior distributions incorporates epistemic uncertainty, producing an unconditional probability of trial success.

1: **Inputs:** Number of iterations N; prior distributions for model parameters; control and treatment data generation models; critical value c_α under H_0
2: **for** $i = 1$ **to** N **do**
3: Sample parameter values from their prior distributions
4: Simulate control and treatment group data conditional on sampled parameters
5: Compute test statistic t_i
6: Set $U_i = \mathbb{1}(t_i > c_\alpha)$ ▷ 1 if H_0 rejected, 0 otherwise
7: **end for**
8: Estimate assurance, where R denotes the event of rejecting H_0:

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

2.5.1 Example: The Moxonidine Trial

To illustrate simulation-based assurance, we consider an example inspired by a clinical trial conducted between 2002 and 2006 (Bolliger et al., 2007). Myocardial ischaemia is a major cause of postoperative morbidity and mortality following coronary artery bypass surgery (Sprung et al., 2000). In the trial, patients were randomised 1:1 to receive either moxonidine or a control treatment. Medication was administered from the morning of surgery until four days postoperatively, with personnel blinded to allocation.

The primary endpoint was the incidence of a clinically significant rise ($\geq$2 μg/L) in cardiac troponin I (cTnI) concentration within seven days after surgery. A previous study reported a 45% incidence in the control group, and investigators hypothesised that moxonidine would reduce this to 30%.

For each arm, let r_i denote the number of patients with an event out of n_i participants ($i = c$ for control, $i = t$ for treatment). The observed proportions are $\hat{\theta}_i = r_i/n_i$, and the hypothesis test is formulated as:

$$H_0 : \theta_e \geq \theta_c \quad \text{vs.} \quad H_1 : \theta_e < \theta_c.$$

Assuming a one-sided Type I error rate of $\alpha = 0.025$ and target power of 80% ($\beta = 0.2$), Algorithm 3 is used to determine the required sample size.

Algorithm 3 Power simulation for the Moxonidine trial. Repeated sampling under fixed rates estimates the probability of rejecting H_0 when the true reduction in event rate is 15%.

1: **Inputs:** Number of iterations N; per-arm sample size n; event probabilities θ_c, θ_e under H_1; critical value z_α
2: **for** $i = 1$ **to** N **do**
3: Sample $r_C \sim \text{Binomial}(n, \theta_c)$ and $r_T \sim \text{Binomial}(n, \theta_e)$
4: Compute sample proportions $\hat{\theta}_C = r_C/n$ and $\hat{\theta}_T = r_T/n$
5: Compute pooled proportion $\hat{\theta}_P = (r_C + r_T)/(2n)$
6: Compute test statistic $z_i = (\hat{\theta}_C - \hat{\theta}_T)/\sqrt{2\hat{\theta}_P(1 - \hat{\theta}_P)/n}$
7: Set $U_i = \mathbb{1}(z_i > z_\alpha)$ ▷ 1 if H_0 rejected, 0 otherwise
8: **end for**
9: Estimate power, where R denotes the event of rejecting H_0:

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

This simulation yields an estimated sample size of approximately 162 patients per arm to achieve 80% power.

To compute *assurance*, prior distributions must be specified for both θ_c and the treatment effect. The prior mean for θ_c is 0.45, with a standard deviation of 0.1. Fitting this uncertainty to a Beta distribution gives $\theta_c \sim \text{Beta}(10.7, 13.1)$, a natural choice for a probability parameter bounded on $[0, 1]$.

Rather than modelling θ_e directly, it is convenient to model the treatment effect $\rho = \theta_c - \theta_e$, recognising that θ_c and θ_e are not independent. Based on expert opinion, clinicians anticipate a mean reduction of 15% in event rate, modelled as:

$$\rho \sim \mathcal{N}(0.15, \nu),$$

where ν quantifies uncertainty in the expected effect.

The full simulation algorithm for assurance is shown in Algorithm 4.

Algorithm 4 Assurance simulation for the Moxonidine trial. Sampling from prior distributions for θ_c and ρ quantifies epistemic uncertainty and yields the unconditional probability of rejecting H_0.

Inputs: Number of iterations N; per-arm sample size n; prior distributions $\pi(\theta_c)$ and $\pi(\rho)$; critical value z_α
for $i = 1$ **to** N **do**
    Sample $\theta_{c,i} \sim \pi(\theta_c)$ and $\rho_i \sim \pi(\rho)$
    Set $\theta_{t,i} = \theta_{c,i} - \rho_i$
    Sample $r_C \sim \text{Binomial}(n, \theta_{c,i})$ and $r_T \sim \text{Binomial}(n, \theta_{t,i})$
    Compute sample proportions $\hat{\theta}_C = r_C/n$ and $\hat{\theta}_T = r_T/n$
    Compute pooled proportion $\hat{\theta}_P = (r_C + r_T)/(2n)$
    Compute test statistic $z_i = (\hat{\theta}_C - \hat{\theta}_T)/\sqrt{2\hat{\theta}_P(1 - \hat{\theta}_P)/n}$
    Set $U_i = \mathbb{1}(z_i > z_\alpha)$ ▷ 1 if H_0 rejected, 0 otherwise
end for
Estimate assurance, where R denotes the event of rejecting H_0:

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

To investigate the effect of different levels of prior uncertainty, three scenarios are considered:

- **Scenario 1:** Highly informative prior centred on a 15% effect, $\rho \sim \mathcal{N}(0.15, 0.0001)$.
- **Scenario 2:** Moderately informative prior with greater uncertainty, $\rho \sim \mathcal{N}(0.15, 0.01)$.

- **Scenario 3:** Moderately informative prior centred on a smaller effect, $\rho \sim \mathcal{N}(0.10, 0.01)$.

Figure 2.3 compares assurance across these scenarios alongside the frequentist power curve. Under Scenario 1, where prior uncertainty is minimal, assurance is effectively equivalent to power. In Scenario 2, wider prior uncertainty reduces the assurance modestly. In Scenario 3, the lower prior mean effect further decreases assurance. These results demonstrate that both the mean and the precision of prior beliefs strongly influence the estimated probability of success. This underscores the importance of careful choice of prior distributions in practical applications of assurance.

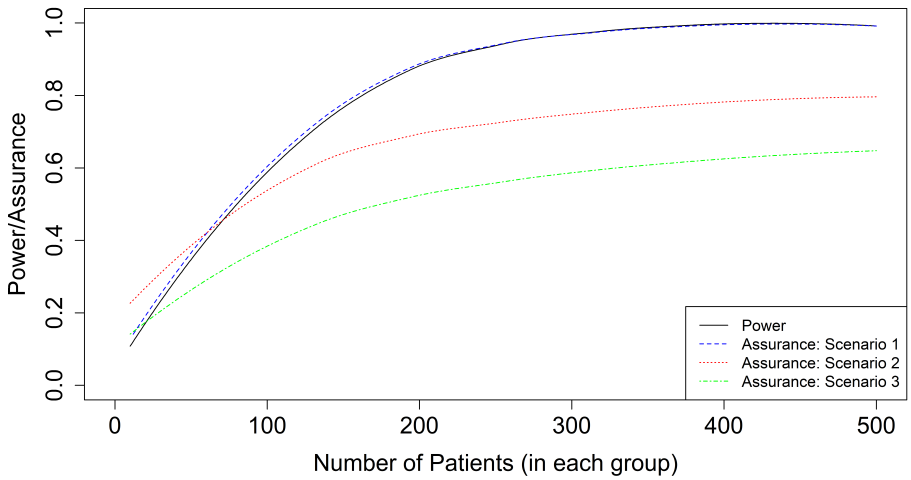


Figure 2.3: *Power and assurance for the Moxonidine trial under different prior distributions for ρ. Stronger (more concentrated) priors yield assurance values*

closer to nominal power, while weaker or pessimistic priors produce lower assurance, highlighting the sensitivity of assurance to prior specification.

Overall, simulation-based methods provide a practical and general framework for calculating assurance in settings where analytical solutions are unavailable. They allow flexible modelling of complex data-generating processes and prior structures, making them particularly valuable for modern clinical trial designs where uncertainty must be explicitly quantified and propagated.

2.6 Development of Assurance Methods

The concept of assurance was introduced by Spiegelhalter and Freedman, 1986 as a means of integrating uncertainty about treatment effects directly into the process of trial design. Since then, the framework has undergone substantial development, spanning analytic extensions, methodological generalisations, and increasingly sophisticated applied implementations. This section summarises key advances that have shaped assurance into a versatile tool for quantitative decision-making in drug development.

Early methodological work concentrated on settings with normally distributed endpoints. O'Hagan, Stevens, et al., 2005 formalised the term assurance and provided general expressions for continuous and binary outcomes, establishing a foundational Bayesian predictive structure. Subsequent refinements addressed more realistic modelling assumptions: for example, Alhussain and Oakley, 2020 derived assurance calculations allowing for an unknown variance, thereby accommodating designs in which variability must be estimated rather than treated as fixed. These contributions grounded assurance firmly within standard continuous-outcome trial planning workflows.

Assurance was then extended to a broader class of data types. Spiegelhalter, Freedman, and Parmar, 1994 developed predictive success probabilities for survival outcomes under proportional hazards assumptions, extending the framework to time-to-event settings. Building on these ideas, Ren and Oakley, 2014 proposed general methods applicable to both parametric and non-parametric survival models, enabling implementation across a wide range of clinical scenarios.

More recently, attention has turned to the dynamical use of predictive success measures across development programmes. Temple and Robertson, 2021 introduced the concept of conditional assurance, quantifying how the probability of success evolves as data accumulate across sequential phases, such as Phase IIa–IIb–III. In parallel, Grieve, 2024 developed assurance calculations tailored to group sequential designs, allowing predictive probabilities to be updated at interim analyses while preserving Type I error control. These contributions highlight the increasing alignment between predictive success metrics and adaptive trial structures.

A substantive strand of recent research has focused on extending assurance to complex trial architectures that involve multiple sources of uncertainty. For example, Wilson, 2023 developed assurance methodology for cluster-randomised trials, demonstrating how uncertainty in parameters such as the intracluster correlation coefficient and cluster sizes can be integrated coherently at the design stage. This work illustrates how the assurance framework can be adapted to settings in which the data-generating structure itself introduces additional layers of variability. Complementing this, Williamson et al., 2025 proposed hybrid approaches that combine assurance with classical power calculations, particularly for cluster-based designs.

Alongside these methodological developments, the application of assurance in industrial settings has grown substantially. Case studies across multiple therapeutic areas, such as those reported by Sabin et al., 2014, Walley et al., 2015, Crisp et al., 2018, Hampson, Holzhauer, et al., 2022, and Hampson, Bornkamp, et al., 2022, demonstrate its role in informing "go/no-go" decisions, quantifying uncertainty in key assumptions, and evaluating strategic options under realistic variability. Examples such as Jiang, 2011 and Hong and Shi, 2012 illustrate how assurance has been used to assess robustness to uncertainties in treatment effect, recruitment patterns, and other operational characteristics. These applied contributions reflect the increasing recognition of assurance as a practical decision-support tool rather than a purely theoretical construct.

In practice, sample size would still typically be determined via a conventional power calculation, with assurance then evaluated at the resulting n to provide a more realistic assessment of the probability of trial success under prior uncertainty. Uncertainty in this evaluation can be communicated by plotting assurance as a function of sample size across a range of plausible effect sizes, giving the trial designer a transparent picture of how sensitive the chosen n is to prior assumptions.

Methodological research continues to evolve towards fully flexible, simulation-based assurance frameworks capable of handling multi-parameter uncertainty and adaptive decision rules. For instance, Wang, Fu, et al., 2013 explored simulation-based methods for assurance in adaptive designs, demonstrating how predictive success metrics can be embedded within designs featuring mid-course decisions. More integrated frameworks, such as those from Hampson, Holzhauer, et al., 2022 and Hampson, Bornkamp, et al., 2022, show how assurance can be incorporated across multiple stages of development, facilitating portfolio-level decision-making.

In summary, the development of assurance methodology reflects a progression from simple analytic formulations to generalisable, simulation-driven approaches that accommodate complex endpoints, adaptive structures, and multi-layered uncertainty. This evolution has established assurance as a central tool linking statistical modelling, decision analysis, and practical considerations in clinical development. Because the usefulness of assurance depends critically on the specification of prior uncertainty, the next chapter examines methods for constructing and eliciting such priors in a transparent and rigorous manner.

2.7 Prior Distribution

A central element in the calculation of assurance is the specification of the prior distribution, $\pi(\theta)$, for the unknown treatment effect θ. This prior encapsulates the current state of knowledge or belief about the likely magnitude of the treatment effect before

data from the planned study are observed. The quality and relevance of this prior are therefore critical, as it directly determines the degree to which epistemic uncertainty is propagated through the assurance calculation.

Broadly, two approaches exist for defining $\pi(\theta)$: one based on the synthesis of relevant historical data, and the other grounded in structured expert judgement. The former provides a data-driven and reproducible mechanism for prior construction, whereas the latter leverages domain expertise in contexts where empirical evidence is limited or only partially applicable.

2.7.1 Data-driven priors

A variety of formal methods have been proposed to derive priors from historical or external data, including the *power prior* (Ibrahim and Chen, 2000), the *commensurate prior* (Hobbs et al., 2011), and the *meta-analytic-predictive (MAP) prior* (Neuenschwander et al., 2010; Schmidli et al., 2014). These approaches were primarily developed in the context of Bayesian borrowing, where external data are incorporated into current analyses to improve precision or reduce required sample size. The robust MAP framework of Schmidli et al., 2014, for example, allows partial borrowing while safeguarding against potential conflicts between historical and current data. Similarly, Viele et al., 2014 provide a structured approach for selecting borrowing strategies according to the degree of compatibility between the data sources.

In principle, the posterior distribution from a preceding study, such as a Phase II trial updated from a weakly informative prior, can serve as $\pi(\theta)$ for the assurance calculation. This approach provides a transparent link between stages of drug development and ensures coherence across analyses. However, it also presents limitations. Differences in design between early and confirmatory trials, including variations in patient population, endpoint definition, or dosing regimen, can undermine the validity of such direct extrapolation. Thus, while historical borrowing offers an attractive starting point, it requires careful consideration of the context and quality of the available evidence.

2.7.2 Expert-elicited priors

When historical data are sparse, heterogeneous, or only partially relevant, expert elicitation provides an alternative route for constructing $\pi(\theta)$ (Chaloner and Rhame, 2001; Dolan et al., 1986). In this framework, subject-matter experts, typically clinicians, statisticians, and researchers familiar with the therapeutic area, quantify their beliefs about key parameters after reviewing the available evidence. The process is often sup-

ported by structured protocols that help translate qualitative beliefs into quantitative prior distributions.

Elicitation is particularly valuable in rare diseases or emerging therapeutic areas, where the evidence base may comprise a mixture of small trials, observational studies, registries, or case series. These sources may not lend themselves to formal statistical synthesis, yet they still inform expert judgement. Structured elicitation thus provides a principled mechanism for integrating such diverse information into a coherent prior (Hampson, Whitehead, et al., 2015). Importantly, it enables the explicit representation of uncertainty and disagreement among experts, features that are often obscured in purely data-driven analyses.

Furthermore, expert elicitation serves a crucial role in bridging early- and late-phase studies. Differences in patient populations, endpoints, or treatment regimens between Phase II and Phase III trials can limit the direct transferability of empirical findings. In such cases, elicited priors provide a means of adjusting expectations based on expert understanding of biological plausibility, treatment mechanisms, or contextual evidence (Holzhauer et al., 2022). By incorporating this judgement transparently, assurance calculations can more accurately reflect realistic prior uncertainty about the treatment's true effect.

2.7.3 Summary

In summary, both historical-data-based and expert-elicited priors provide valid foundations for assurance, but each carries distinct strengths and limitations. Data-driven approaches emphasise reproducibility and objectivity, while elicitation allows flexibility and contextual insight when evidence is limited. In practice, the most defensible priors often result from combining these approaches, using available data to inform, calibrate, or constrain expert belief. The next chapter (Chapter 3) explores in detail the methodology and practical considerations involved in eliciting expert opinions to construct priors suitable for assurance-based trial design.

2.8 Summary

This chapter introduced assurance methods as a Bayesian alternative to traditional power calculations for clinical trial design. We began by revisiting the classical frequentist definition of power, highlighting its dependence on a fixed, assumed treatment effect. By integrating prior uncertainty about the true treatment effect through the concept of assurance, we obtained a more realistic and probabilistically coherent

measure of a trial's likelihood of success.

Analytic examples, such as the Normal–Normal model, demonstrated how assurance can be expressed in closed form when conjugate priors are used. These examples illustrated key properties of assurance, including its asymptotic upper bound and the interpretation of *normalised assurance* as a relative efficiency measure. We then extended the discussion to simulation-based approaches, which are required in more realistic settings where closed-form solutions are unavailable. The moxonidine case study illustrated how Monte Carlo simulation can be used to evaluate assurance under different prior beliefs, emphasising the influence of both the strength and the uncertainty of those priors on the resulting probability of trial success.

The development of assurance methods was then situated in its historical and applied context, tracing its evolution from early theoretical foundations (Spiegelhalter and Freedman, 1986; O'Hagan, Stevens, et al., 2005) to its modern use in complex and adaptive trial designs (Grieve, 2024; Hampson, Bornkamp, et al., 2022). Particular attention was given to the increasing adoption of assurance within industry, where it now forms an integral part of quantitative decision-making and risk assessment frameworks.

Finally, we discussed the crucial role of the prior distribution in assurance calculations. The prior serves as the bridge between existing knowledge and future uncertainty, and its careful specification determines the credibility and interpretability of the resulting assurance. We reviewed both data-driven and expert-elicited approaches to constructing priors, highlighting their respective advantages and limitations.

Overall, assurance provides a Bayesian framework for quantifying the probability of trial success while accounting for prior uncertainty. It extends beyond traditional power calculations by embedding design evaluation within a predictive, decision-oriented context. The next chapter builds on this foundation by focusing on methods for prior elicitation, formal techniques for capturing and quantifying expert belief, thereby addressing one of the most practically important aspects of implementing assurance in real-world trial design.

Chapter 3

Expert Elicitation

3.1 Introduction

The calculation of assurance in clinical trials requires the specification of a prior distribution for one or more design parameters, most commonly the treatment effect. This prior formalises existing knowledge and uncertainty before the trial begins, and plays a central role in probability-of-success evaluation. While empirical evidence from historical studies, meta-analyses, or mechanistic models can inform such priors, these sources are often limited, heterogeneous, or of uncertain relevance, particularly for novel therapies, rare diseases, or early-phase development. In these settings, the most transparent and defensible approach to prior specification is the use of structured expert judgement, or *expert elicitation*. When used appropriately, expert elicitation provides a coherent means of incorporating knowledge that cannot readily be obtained from data alone. However, poorly designed or uncritical elicitation exercises risk producing overconfident or misleading priors, underscoring the need for structured protocols and transparency (Morgan, 2014).

Expert elicitation is a disciplined methodology for capturing the beliefs of domain specialists about uncertain quantities and expressing those beliefs as probability distributions. It provides a principled mechanism for incorporating qualitative clinical insight into formal statistical frameworks, ensuring that the assumptions underpinning trial design are explicit, reproducible, and open to scrutiny. Within assurance-based design, elicited priors enable uncertainty about treatment effects and other design parameters to be represented coherently, supporting probability-of-success calculations and operating-characteristic evaluation when empirical data are limited or uninformative.

The practice of expert elicitation draws on decades of development in statistics, deci-

sion theory, risk analysis, and psychology. Numerous structured frameworks have been proposed to guide the process, reduce cognitive bias, and improve the validity of expert judgements. Although the broader literature is extensive, this chapter focuses on the methods most relevant to constructing priors for clinical trial design and assurance calculations. Topics include the selection of experts, the formulation of elicitation questions, the fitting of probability distributions to expert judgements, and considerations surrounding aggregation, calibration, and communication of uncertainty.

By synthesising methodological principles with practical guidance, this chapter establishes the foundation for the delay-specific elicitation framework developed in Chapter 4. It highlights that, while the mathematics of assurance is straightforward once a prior is available, the process of obtaining that prior is a complex and fundamentally human activity requiring structured protocols, careful facilitation, and rigorous statistical interpretation.

3.2 Aims of the Chapter

The specific objectives of this chapter are as follows:

1. To explain the role and importance of expert knowledge in informing prior distributions, particularly in settings where data are sparse, uncertain, or not directly applicable to the target population.

2. To describe key methodological frameworks for expert elicitation, including both structured and semi-structured protocols, and to highlight their theoretical underpinnings.

3. To present practical guidance for conducting elicitation exercises, encompassing expert selection, question design, facilitation, and approaches to mitigate common cognitive biases.

4. To discuss the methodological and practical challenges associated with elicitation, including issues of reproducibility, calibration, and communication of uncertainty.

5. To demonstrate how elicited priors can be integrated into assurance calculations and Bayesian trial designs, thereby supporting quantitative decision-making in clinical development.

3.3 Methods for Eliciting Univariate Distributions

The goal of expert elicitation is to obtain a probability distribution that accurately reflects an expert's belief about an uncertain parameter, denoted here by θ (for example, a treatment effect or event probability). Most domain experts are not accustomed to expressing beliefs as probability distributions, and even those with statistical training often find it challenging to specify a full distribution directly. To address this, several elicitation methods have been developed to help experts express their uncertainty in intuitive ways that can then be translated into formal statistical distributions.

This section outlines the most widely used elicitation methods, drawing primarily on the guidance of O'Hagan, Buck, et al., 2006, alongside more recent critical reviews of elicitation approaches that highlight the diversity of available methods and their respective strengths and limitations (Falconer et al., 2022). For clarity, the discussion begins by considering the elicitation of beliefs from a single expert. In practice, however, elicitation exercises commonly involve multiple experts, whose individual judgements must be combined or synthesised; these issues are addressed in Section 3.5.

3.3.1 Quantile-Based Methods

Quantile-based elicitation is one of the simplest and most widely used approaches. Experts are asked to specify values of θ corresponding to particular quantiles of their belief distribution. Common variants used in practice include:

- The *quartile method*, where the expert provides the 25th percentile, median, and 75th percentile.
- The *tertile method*, where the 33rd and 66th percentiles are elicited instead.
- Eliciting the 5th, 50th, and 95th percentiles, providing a broader characterisation of the tails of the expert's belief distribution.

The elicited quantiles are then used to fit a suitable parametric distribution representing the expert's uncertainty (see Section 3.3.4), although non-parametric representations are also possible. Depending on the nature of the parameter, commonly used families include the Normal, Log-Normal, Beta, Gamma, and Student-t distributions, among others.

Quantile-based methods are cognitively straightforward and impose minimal burden on participants. However, some experts may find it difficult to reason about exact

percentiles, particularly in the tails of the distribution, and the resulting fitted distribution may be sensitive to the chosen parametric form, especially when only a small number of quantiles are elicited.

3.3.2 Probability-Based Methods

In probability-based methods, experts are asked to assign probabilities to predefined intervals or threshold values of θ. For example, they may be asked:

- "What is your probability that θ lies between 10 and 15?"
- "What is your probability that θ is less than 0.4?"

This approach is often more intuitive for experts who prefer reasoning in terms of likelihoods rather than percentiles. However, care must be taken to avoid introducing anchoring bias through the presentation of fixed thresholds (see Section 3.7). Randomising or rotating the order of threshold questions, or allowing experts to define their own intervals, can help mitigate this risk. Empirical comparisons of elicitation formats indicate that probability-based judgements can exhibit greater variability and monotonicity violations than fixed-quantile approaches (Abbas et al., 2008), underscoring the need for careful design of threshold-based questions.

3.3.3 Histogram- or Roulette-Based Elicitation

Histogram-based methods provide a visual and tactile way for experts to express their uncertainty. The *Trial Roulette Method* (Gore, 1987; Johnson et al., 2010) is one of the best-known examples. Experts are asked to distribute a fixed number of "chips" or tokens across a series of bins representing intervals over a plausible range for θ. The proportion of chips allocated to each bin represents the subjective probability that θ lies within that interval.

In practice, the range of θ is divided into 8–12 bins, and the expert is typically given around 20 chips (each representing 5% probability). This provides a good balance between precision and cognitive simplicity. Too many chips may exceed an expert's capacity to make fine distinctions, while too few may fail to capture the shape of their uncertainty.

Empirical studies suggest that experts often find this method intuitive and engaging, as it resembles direct manipulation of probabilities. However, facilitators must remind

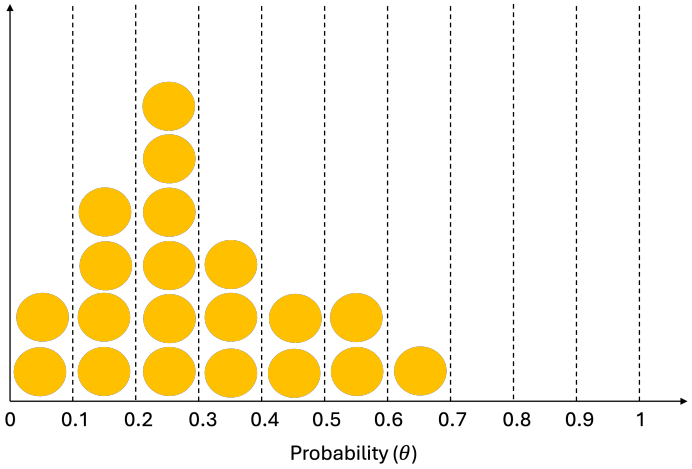


Figure 3.1: *Example of the 'Roulette' method for eliciting beliefs about a parameter θ. The expert allocates probability mass across predefined bins using a fixed number of chips.*

participants that their distributions need not be symmetrical or visually neat, since the goal is to express genuine uncertainty rather than aesthetic balance.

Feedback should always be provided after initial elicitation. For example, facilitators might show the expert implied probabilities from their histogram (e.g., a 25% chance that $\theta > 0.4$ and 10% that $\theta < 0.1$) to confirm that the representation is consistent with their beliefs.

3.3.4 Fitting Probability Distributions

Once elicited data (such as quantiles or histogram bins) have been obtained, the next step is to fit a continuous probability distribution that approximates the expert's beliefs. While the elicited histogram itself could be used, parametric distributions are usually preferred for analytical convenience and ease of simulation.

Common choices include the Normal, Beta, and Gamma distributions, depending on the domain and the support of θ. Parameters are often estimated via least squares—minimising the difference between the elicited and fitted quantiles or probabilities.

Figure 3.2 shows an example where a Beta(2.04, 4.89) distribution provides a good fit to the histogram elicited via the roulette method (Figure 3.1).

Although nonparametric approaches are available (Oakley and O'Hagan, 2007; Gosling et al., 2007), they are less common in practice due to computational complexity and

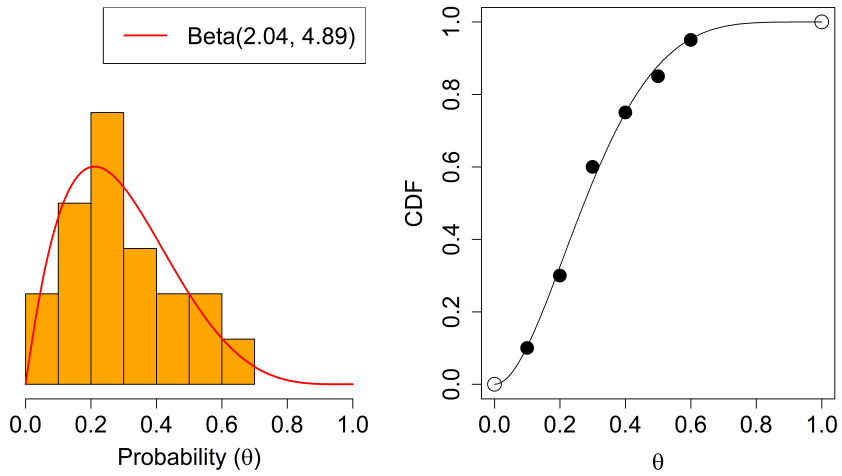


Figure 3.2: *Example from Figure 3.1. a) The best-fitting Beta distribution is overlaid (Beta(2.04, 4.89)); b) The elicited quantile points with the fitted Beta CDF overlaid. The fitted distribution is reviewed with the expert to confirm adequacy of fit.*

challenges in implementation. After fitting, experts should be shown diagnostic plots or summaries to verify that the fitted distribution accurately represents their beliefs. Sensitivity analyses may also be performed to assess the robustness of downstream results to the chosen distributional form.

3.4 Eliciting Multiple Uncertain Quantities

In many practical settings, elicitation involves more than one uncertain parameter. For example, when designing a clinical trial, experts may need to provide beliefs about both the control response rate and the treatment effect, or about several correlated model parameters such as baseline hazard and treatment hazard ratio. In such cases, it becomes important to consider whether these quantities can be assumed to be statistically independent, or whether dependencies must be explicitly captured.

Formally, two quantities θ_1 and θ_2 are independent if knowledge of one does not change beliefs about the other. If an expert's belief about θ_1 would be revised upon learning the value of θ_2, then these quantities are dependent, and a joint distribution must be elicited. Dependencies are common in clinical and pharmacological contexts, where parameters often reflect related biological or mechanistic processes.

If independence can be reasonably assumed, quantities can be elicited separately using standard univariate methods described in Section 3.3. However, when dependencies are expected, additional elicitation is needed to characterise both the direction and magnitude of correlation. This can be a challenging task: dependencies are inherently abstract, and experts often find it difficult to express beliefs about correlations numerically or verbally.

A variety of approaches have been proposed for eliciting dependence structures. A common and pragmatic strategy is to *structure* the problem by reparameterising quantities in ways that make independence more plausible. For instance, instead of eliciting the treatment effect and control rate directly, one might elicit the control rate and the treatment–control difference. If the difference is believed to be unrelated to the baseline level, independence can then be assumed. This strategy, described by Garthwaite et al., 2005 and Daneshkhah and Oakley, 2010, can often simplify multivariate elicitation considerably.

When structuring is not feasible, several modelling approaches are available to represent expert beliefs about dependency. The use of copulas, particularly Gaussian copulas, provides a flexible way to model correlation between continuous quantities without constraining marginal distributions. For proportions that must sum to one, such as probabilities of mutually exclusive outcomes, Dirichlet distributions can be employed to represent joint beliefs while preserving logical constraints (Zapata-Vázquez et al., 2014). Both approaches have been implemented in extensions to the SHELF framework (Oakley and O'Hagan, 2025), allowing experts to visualise and validate the implied relationships among quantities.

Direct elicitation of correlations remains difficult, as experts tend to underestimate or misjudge the strength of relationships between uncertain quantities. To address this, graphical tools and iterative feedback can be useful. For example, showing implied scatterplots or probability contours can help experts assess whether the dependence encoded in a fitted joint distribution matches their intuitive understanding.

For a comprehensive discussion of techniques for eliciting joint distributions and dependence structures, see Werner, Bedford, et al., 2017; Werner, Hanea, et al., 2017, who provide a taxonomy of approaches for continuous and discrete variables. More recent applied work by Holzhauer et al., 2022 illustrates how structured elicitation of correlated parameters can be implemented in pharmaceutical contexts, particularly for survival and dose–response modelling.

Overall, while multivariate elicitation presents additional complexity, careful problem structuring, graphical validation, and iterative engagement with experts can yield joint priors that are both realistic and practically usable within assurance-based trial designs.

3.5 Combining Judgements from Multiple Experts

Elicitation exercises in clinical trial design often involve more than one expert, particularly when the parameters of interest span multiple domains of expertise, such as clinical outcomes, pharmacology, and statistical modelling. Combining judgements from several experts can improve robustness by integrating diverse perspectives, reducing individual bias, and broadening the evidence base on which priors are constructed. However, experts rarely agree perfectly, and synthesising heterogeneous opinions into a single coherent probability distribution is one of the central challenges of structured elicitation.

Two main paradigms exist for combining expert judgements: *mathematical aggregation* and *behavioural aggregation*. These represent fundamentally different philosophies. The first treats expert distributions as statistical inputs to be combined using a formal rule, while the second treats experts as collaborators engaged in a structured process of deliberation to reach a shared view. The choice between these paradigms depends on the study objectives, logistical constraints, and the desired degree of transparency and interaction.

3.5.1 Mathematical Aggregation

Mathematical aggregation synthesises expert opinions without direct interaction among participants. Each expert provides an individual probability distribution $f_i(\theta)$, which are then combined into a single consensus distribution $f(\theta)$ using a specified pooling rule. Most pooling rules implicitly assume that expert judgements are conditionally independent; however, empirical analyses of multi-expert elicitation studies suggest that between-expert dependence is common and can materially affect the properties of aggregated distributions (Wilson, 2017).

The most widely used approaches are the *linear pool* and the *logarithmic pool* (Clemen and Winkler, 1999; O'Hagan, Buck, et al., 2006).

The linear pool combines individual expert distributions through a weighted average,

$$f(\theta) = \sum_{i=1}^{n} w_i f_i(\theta),$$

where w_i denotes the weight assigned to expert i with $\sum_i w_i = 1$. This approach is intuitive, computationally straightforward, and preserves the full range of uncertainty expressed by the experts. However, when expert judgements differ substantially, the resulting distribution can be overly diffuse, reflecting disagreement rather than synthesising it.

The logarithmic pool instead combines distributions multiplicatively,

$$f(\theta) \propto \prod_{i=1}^{n} f_i(\theta)^{w_i}.$$

This method yields a consensus distribution that places greatest mass where expert opinions overlap. It typically produces a narrower pooled distribution than the linear pool, reflecting greater consensus among experts, though this comes at the cost of down-weighting the tails of individual distributions — which may be problematic if extreme values reflect genuine uncertainty rather than elicitation noise.

Figure 3.3 illustrates these differences for two experts pooled with equal weights. Under the linear pool, the combined distribution retains support across the union of the experts' beliefs, resulting in a wider and more diffuse distribution that reflects between-expert disagreement. Under the logarithmic pool, the combined distribution concentrates probability mass in regions where expert opinions overlap, yielding a more concentrated consensus distribution. The choice between pooling methods should therefore reflect whether the objective is to preserve the full extent of expert uncertainty or to emphasise areas of agreement across experts.

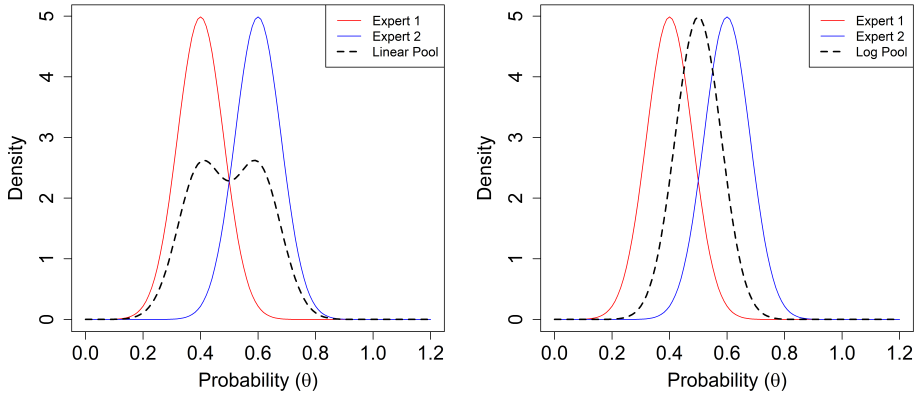


Figure 3.3: *Comparison of the linear and logarithmic pooling methods applied to two expert distributions with equal weights. The linear pool retains the combined support of both experts, whereas the logarithmic pool concentrates mass in regions where the experts agree.*

A key practical consideration is the assignment of weights w_i. Equal weighting is simple and often used when all experts have similar standing. However, when there is reason to believe that certain experts are more reliable, weights can be derived from performance-based metrics. Cooke's Classical Method (Cooke, 1991) uses calibration questions to evaluate expert accuracy and informativeness. These scores are then used to generate

objective weights. Although effective, this approach can be resource-intensive and depends on the availability of suitable calibration quantities.

More advanced approaches, such as the *supra-Bayesian framework* (Jacobs, 1995), treat expert opinions as data, combining them through Bayes' theorem with a decision-maker's prior. This provides a fully Bayesian rationale for aggregation but requires complex hierarchical modelling and substantial computational effort. In practice, linear or logarithmic pooling remains the most common strategy in applied work due to its simplicity and interpretability.

3.5.2 Behavioural Aggregation

Behavioural aggregation combines expert judgements through structured discussion and deliberation rather than through purely mathematical pooling. Experts interact directly (either in person or virtually) under the guidance of a trained facilitator. The aim is to develop a shared understanding of the problem, clarify reasoning, and, where possible, agree on a collective probability distribution that captures the group's consensus. The Sheffield Elicitation Framework (SHELF; Section 3.6.1) is a prominent example of behavioural aggregation. Experts present and justify their beliefs individually before engaging in facilitated discussion to reach a consensus distribution known as the *Rational Impartial Observer (RIO)* prior. This represents the belief that a neutral, well-informed observer would hold after hearing all arguments. The RIO prior is not a simple average — it reflects deliberation, justification, and reconciliation of differing views. Figure 3.4 is a fictional example, illustrating how the RIO prior can balance divergent individual beliefs to form a coherent group distribution. Behavioural aggregation relies heavily on the skill of the facilitator, the clarity of the elicitation questions, and the willingness of experts to engage in open discussion. When well-executed, it can yield distributions that are more coherent, transparent, and psychologically satisfying to participants than those produced by mathematical pooling.

3.5.3 Hybrid Approaches

Some aggregation frameworks combine elements of both behavioural and mathematical aggregation, and are therefore more accurately described as hybrid approaches. Rather than relying solely on open facilitated discussion or purely on statistical pooling, these methods incorporate structured interaction or iterative feedback to guide experts towards a collective judgement. The Delphi method (Dalkey and Helmer, 1963) maintains anonymity between experts, using multiple iterative rounds of anonymous feedback and revision to promote convergence. While this approach limits collaborative reasoning, it helps avoid dominance or conformity effects that can arise in face-to-face settings,

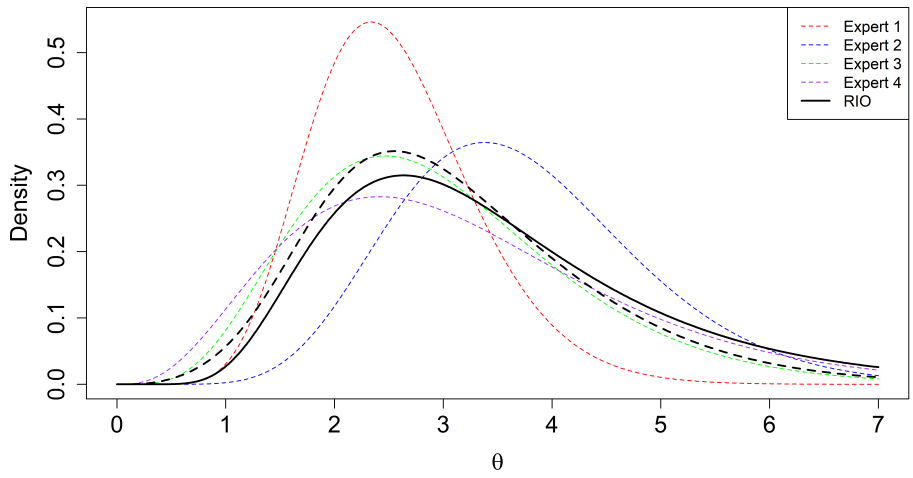


Figure 3.4: *Example of a RIO prior derived from four experts with differing individual beliefs.*

making it particularly useful when power dynamics between experts are a concern. The IDEA protocol (Hemming et al., 2018) similarly combines structured individual elicitation with a subsequent aggregation step, incorporating elements of both anonymous feedback and mathematical pooling. By separating individual judgement from group reconciliation, IDEA aims to balance the benefits of independent reasoning with the coherence gains of structured deliberation.

3.5.4 Comparing and Selecting Aggregation Approaches

The choice between aggregation approaches has substantive implications for the form and credibility of the elicited prior. Mathematical pooling offers reproducibility, ease of implementation, and analytical tractability, features that are particularly attractive when priors are used within formal decision models such as assurance or probability-of-success analyses. Behavioural aggregation, by contrast, emphasises structured discussion, conditional reasoning, and shared understanding among experts, often producing consensus distributions that stakeholders regard as more transparent and defensible.

Williams et al., 2021 conducted the first systematic empirical study evaluating the differences between Cooke's Classical Method and the SHELF behavioural framework. Their findings highlight that aggregation choice can meaningfully influence both the shape and location of the resulting distributions. In particular, performance-weighted pooling under the Classical Method tends to produce more concentrated distributions when calibration scores vary substantially across experts, whereas SHELF's consensus-

based approach often leads to broader or more conservative priors reflecting the group's collective reasoning. The study also underscores practical considerations: the Classical Method offers objectivity through explicit weighting, while SHELF facilitates richer qualitative justification and fosters convergence on quantities for which direct probabilistic reasoning is challenging. These contrasts are especially relevant when the elicited priors feed into downstream decision processes, such as assurance calculations, where differences in tail behaviour or central tendency may materially affect predicted trial success.

Selecting an appropriate aggregation strategy requires aligning methodological properties with the aims of the elicitation. Mathematical pooling may be preferable when expert independence is well supported or when reproducibility is a key priority. Behavioural aggregation is often more suitable for complex scientific judgements involving interdisciplinary expertise or situations where stakeholder consensus is critical. Hybrid approaches that combine elements of both can offer a pragmatic compromise. For example, behavioural elicitation may be used to develop individual distributions, followed by mathematical pooling of the post-discussion judgements. The suitability of such approaches should be assessed on a case-by-case basis.

Ultimately, aggregation is not solely a technical operation but a process of constructing and justifying a shared representation of uncertainty. A considered choice of approach enhances both the epistemic integrity and practical utility of the resulting priors, ensuring that they offer a defensible basis for subsequent decision-making in clinical trial design.

3.6 Protocols

In expert elicitation, a protocol refers to a structured and systematic process used to gather, quantify, and document expert judgements in a transparent and reproducible way. The use of formal protocols, often termed structured expert elicitation, has become standard practice across disciplines such as health technology assessment, environmental science, and risk analysis. These frameworks aim to reduce bias, improve consistency, and ensure that expert knowledge is incorporated in a defensible and methodologically sound manner (European Food Safety Authority, 2014).

Several established protocols have been developed to guide elicitation practice, each differing in their philosophy, structure, and intended application. The most commonly used in biomedical research include the Sheffield Elicitation Framework (SHELF), the Delphi method, Cooke's Classical Method, and the IDEA protocol (Investigate, Discuss, Estimate, Aggregate). Each offers a different balance between independence and discussion, between qualitative reasoning and quantitative synthesis. The choice of pro-

tocol depends on practical constraints such as expert availability, time, and whether a consensus or individually weighted aggregation is required (Soares et al., 2024). The following subsections summarise the key features of these major approaches.

3.6.1 Sheffield Elicitation Framework (SHELF)

The Sheffield Elicitation Framework (SHELF) (Oakley and O'Hagan, 2025) is one of the most widely applied protocols in health and risk modelling. It provides a systematic structure for eliciting expert beliefs, typically within a facilitated workshop setting. SHELF is particularly designed for contexts where evidence is limited but transparent and defensible prior distributions are required, such as in health technology assessments or clinical trial design.

A defining feature of SHELF is the construction of a consensus distribution representing the beliefs of a hypothetical Rational Impartial Observer (RIO). This RIO distribution is not simply an average of individual opinions but rather the judgement that an impartial observer might make after considering all experts' reasoning. SHELF workshops usually involve a facilitator, a recorder, and one or more experts. Individual judgements are elicited first (for example, using quantile or roulette methods), followed by group discussion to develop a shared understanding and reach consensus.

SHELF is supported by a dedicated R package (Oakley, 2025), which provides tools for data entry, visualisation, and fitting of parametric distributions (e.g., Normal, Log-Normal, or Beta). The software also allows users to compare consensus-based results with mathematically aggregated alternatives such as linear pooling. Its structured guidance and accessibility have made SHELF a leading choice for elicitation exercises in the pharmaceutical industry and academia.

3.6.2 Probabilistic

Delphi Method

The probabilistic Delphi method (Dalkey and Helmer, 1963) is an iterative, questionnaire-based approach designed to build consensus while preserving expert anonymity. Originally developed for forecasting in defence research, it is now widely used in policy, healthcare, and technology evaluation. In a typical Delphi exercise, experts independently provide quantitative or qualitative judgements, which are summarised and anonymised by a facilitator. Each participant then receives feedback showing how their responses compare with those of others and is invited to revise their judgement in subsequent rounds.

This iterative process continues until the responses converge or a pre-specified criterion is reached. The anonymity of Delphi reduces social pressures such as dominance by vocal individuals or conformity bias, promoting independent reasoning. However, because direct discussion is limited, opportunities for clarifying assumptions or building shared understanding can be constrained. Delphi is most appropriate when experts cannot meet in person, when maintaining independence is vital, or when consensus must be achieved systematically and transparently.

3.6.3 Cooke's Classical Method

Cooke's Classical Method (Cooke, 1991) is a performance-based framework that weights expert judgements according to demonstrated accuracy. Experts first answer a series of calibration questions, quantities from the same domain for which true values are known. Their performance on these questions, in terms of statistical accuracy and informativeness, is used to calculate weights that are then applied when combining their opinions about the target quantities. Recent work on the Classical Method has elaborated on the importance of calibration in ensuring that performance weights accurately reflect an expert's ability to predict real-world outcomes (Cooke and Solomatine, 1992; Colson and Cooke, 2018).

This approach produces a mathematically aggregated distribution that reflects both the expertise and reliability of each participant. The method emphasises independence, with minimal group interaction to avoid bias, and it is supported by software such as EXCALIBUR (Cooke and Solomatine, 1992). Cooke's framework has been applied in diverse fields including nuclear safety, climate science, and epidemiology. Although calibration can be demanding to implement, requiring suitable "calibration" questions and careful design, it remains one of the most rigorous and transparent methods for aggregating expert opinion.

3.6.4 Investigate, Discuss, Estimate, Aggregate (IDEA)

The IDEA protocol (Hemming et al., 2018) aims to balance the strengths of independent judgement with the benefits of structured discussion. Experts first Investigate the problem individually, formulating their own estimates; they then Discuss their reasoning within a facilitated group; subsequently, they Estimate again after considering new perspectives; and finally, their judgements are Aggregated—typically through equal weighting.

IDEA was designed to improve the calibration and accuracy of expert estimates by explicitly incorporating opportunities for reflection and learning. This approach has been

successfully used in ecology, public health, and cost-effectiveness modelling. Compared with purely statistical aggregation (as in Cooke's method), IDEA emphasises dialogue and understanding, helping experts articulate uncertainty in complex or ambiguous contexts.

Overall, the choice of elicitation protocol depends on the balance between available resources, desired level of interaction, and the objectives of the elicitation. In Bayesian clinical trial design, protocols like SHELF and IDEA are most commonly adopted because they combine structure, transparency, and flexibility—making them well suited to developing prior distributions for assurance calculations.

3.7 Challenges of Conducting Prior Elicitation

Although structured elicitation frameworks such as SHELF and IDEA provide detailed guidance, implementing a successful elicitation exercise in practice remains challenging. Difficulties arise at every stage of the process, from selecting suitable experts to interpreting their judgements, and often stem from a combination of technical, logistical, and psychological factors. Recognising and addressing these challenges is essential to ensure that elicited priors are credible, reproducible, and useful for decision-making (European Food Safety Authority, 2014).

3.7.1 Stages of the Elicitation Process

O'Hagan, Buck, et al., 2006 outline a systematic framework for expert elicitation consisting of five stages:

1. **Background and preparation** – defining the objectives of the elicitation and identifying the quantities of interest;
2. **Selecting experts** – identifying individuals with relevant expertise and ensuring diversity of perspective;
3. **Training and motivation** – familiarising experts with probabilistic reasoning, elicitation formats, and common cognitive biases;
4. **Structuring and decomposition** – defining the model parameters, their relationships, and how dependence between quantities will be handled;
5. **Elicitation and feedback** – conducting the sessions, fitting distributions to the expert inputs, and validating results.

While this provides a strong procedural foundation, experience has shown that challenges frequently arise in practice—particularly those related to cognitive bias, communication of uncertainty, and the interpretation of aggregated results.

3.7.2 Cognitive Biases and Heuristics

Perhaps the most pervasive challenges in expert elicitation stem from well-documented cognitive biases. According to the dual-process theory of cognition proposed by Kahneman, 2011, human reasoning operates through two systems:

- **System 1**, which is fast, intuitive, and automatic; and
- **System 2**, which is slower, analytical, and deliberate.

Experts, like all decision-makers, frequently rely on System 1 heuristics, which, while efficient, can produce systematic errors in probabilistic judgement. According to O'Hagan, Buck, et al., 2006, the most common biases encountered in elicitation are:

1. **Anchoring Bias** – Experts may anchor their estimates on an initial reference point (e.g., a published trial result or the facilitator's example) and adjust insufficiently. Anchoring can be particularly problematic when experts are shown early summaries of existing evidence, as these can unduly constrain their responses.
2. **Overconfidence** – Experts tend to underestimate uncertainty, producing distributions that are too narrow. This results in priors that overstate precision, leading to inflated assurance estimates. Encouraging experts to reflect explicitly on worst-case and best-case scenarios can help counteract this tendency.
3. **Availability Bias** – Judgements are influenced by how easily relevant examples come to mind. Salient or recent experiences may disproportionately affect beliefs, even if they are unrepresentative of the broader evidence base.
4. **Group Dynamics** – In group settings, social factors such as dominance by a single expert, conformity pressure, or groupthink can distort the aggregation of opinions. These effects are particularly relevant in behavioural aggregation frameworks such as SHELF and IDEA, where facilitated discussion is integral.

Mitigating these biases requires careful design and facilitation. Experts should receive training on probabilistic thinking and bias awareness before elicitation begins. Questions should be neutrally worded, and quantitative feedback should be provided iteratively to allow experts to reflect on the implications of their judgements.

3.7.3 Training and Facilitation

Training experts prior to elicitation is critical to achieving reliable and interpretable results. Even experienced clinicians or scientists may lack familiarity with expressing beliefs in probabilistic terms. As part of the SHELF framework, an online training course (https://shelf.sites.sheffield.ac.uk/e-learning-course) provides a structured introduction to concepts such as quantiles, credible intervals, and calibration. Training also helps align experts' understanding of uncertainty representation and reinforces the need for internal consistency in their judgements.

The facilitator plays a pivotal role in guiding the elicitation session. They must ensure clarity in communication, maintain neutrality, and balance engagement with efficiency. In behavioural aggregation, this includes moderating discussion to ensure all voices are heard and preventing overdominance by confident individuals. In mathematical aggregation, it includes managing data quality and ensuring that each expert's responses are complete, interpretable, and logically consistent.

3.7.4 Practical and Logistical Challenges

Elicitation exercises can be time-consuming and resource-intensive, especially when multiple experts or parameters are involved. Practical difficulties often include:

- Scheduling and coordinating experts with limited availability;
- Ensuring that elicitation sessions remain focused and do not drift into unstructured discussion;
- Communicating complex statistical quantities in intuitive terms;
- Managing uncertainty about model structure or parameterisation when defining the quantities to be elicited.

Pilot sessions and mock elicitation rounds are often invaluable for refining question design and ensuring smooth implementation.

3.7.5 Validation and Feedback

Validation is a critical but sometimes neglected component of elicitation. Once a probability distribution has been fitted to the elicited inputs, it should be reviewed with the expert to confirm that it accurately reflects their beliefs. This "feedback loop"

allows inconsistencies or misinterpretations to be corrected before the distribution is used in formal analysis. Facilitators should present summaries graphically (e.g., as histograms, density plots, or cumulative distributions) and ask questions such as "Do you agree that this distribution implies a 20% probability that the effect exceeds 0.5?" to check face validity.

In complex or high-stakes applications, validation may extend to comparing elicited priors against emerging data, running sensitivity analyses, or conducting post-hoc calibration assessments to evaluate predictive accuracy over time.

3.7.6 Summary of Common Pitfalls

In summary, the main challenges of prior elicitation arise not from a lack of statistical methods, but from human and procedural factors. Overconfidence, anchoring, and poor facilitation can lead to unrealistic priors; lack of training and validation can undermine credibility; and inadequate documentation can limit transparency. Addressing these challenges requires deliberate effort in the design, conduct, and reporting of elicitation exercises. When managed effectively, elicitation can yield priors that are both statistically coherent and defensible to regulators and decision-makers—enhancing the reliability of assurance calculations and the quality of Bayesian trial designs.

3.8 Elicitation with Assurance Methods

3.8.1 Motivation and Context

In the pharmaceutical industry, the use of elicited priors for assurance calculations has gained considerable traction over the past decade (Best et al., 2020). Early applications were largely confined to academic case studies (Hiance et al., 2009; Kinnersley and Day, 2013), whereas more recent work has demonstrated their value in real-world drug development settings (Dallow et al., 2018; Holzhauer et al., 2022). The appeal of this approach lies in its ability to connect expert clinical understanding of mechanisms of action, patient heterogeneity, and treatment plausibility to formal statistical decision-making at the design stage.

Recent evidence also shows considerable methodological heterogeneity in how elicitation is implemented across clinical trial settings. A systematic mapping study of treatment-effect and borrowing-parameter elicitation identified substantial variation in elicitation targets, protocol choice, aggregation methods, and reporting practice, particularly in early-phase and rare-disease applications (Morgan et al., 2025). These

findings reinforce the need for structured, transparent elicitation procedures when priors are used to inform assurance or probability-of-success calculations.

3.8.2 Practical Implementation

In practice, elicitation-based assurance follows a three-stage workflow:

1. **Define the elicitation target.** Identify the key model parameters that drive the trial outcome—typically the treatment effect or treatment–control difference on the relevant scale (e.g., mean difference, log-odds ratio, or hazard ratio). Ensure these quantities are expressed in clinically interpretable terms, for example by providing benchmark values or contextual reference points to aid expert reasoning.

2. **Conduct the elicitation.** Use a structured framework (such as SHELF or IDEA) to obtain expert beliefs about the plausible range and shape of the target parameter distribution. Where feasible, multiple experts should contribute to ensure diverse perspectives and reduce idiosyncratic bias. Individual beliefs can then be aggregated using either behavioural consensus (e.g., a RIO prior) or mathematical pooling methods.

3. **Compute assurance.** The elicited prior distributions are incorporated into the assurance calculation by integrating over the sampling distribution of the data, as discussed in Chapter 2. Simulation-based approaches are typically required, particularly for complex endpoints or adaptive designs.

An important consideration is the alignment of the elicitation framework with the statistical model used for assurance. For example, in time-to-event studies where assurance is based on a log-hazard ratio, experts may find it easier to express beliefs on more interpretable scales, such as median survival time or absolute risk reduction, before these are transformed to the modelling scale. Such reparameterisation can substantially improve the quality and consistency of elicited inputs.

3.8.3 Case Studies and Applications

Several published case studies illustrate how expert elicitation has been successfully embedded into assurance-based design. Dallow et al., 2018 describe its adoption at GlaxoSmithKline (GSK), where structured elicitation is routinely used to inform priors

for Phase III development programs. Similarly, Holzhauer et al., 2022 report applications at Novartis, where elicitation is used not only for assurance but also to calibrate prior assumptions in Bayesian decision analyses across multiple program stages.

Earlier examples by Hiance et al., 2009 and Kinnersley and Day, 2013 demonstrate how elicited priors can be derived from small panels of clinical experts when historical evidence is limited. More recently, Best et al., 2020 highlight an increasing regulatory acceptance of elicitation-based priors in assurance and Bayesian design evaluation, particularly when accompanied by transparent documentation and validation procedures.

A comprehensive review by Azzolina et al., 2021 summarises the broader literature on elicitation in the design and analysis of clinical trials, noting that while most applications have focused on early-phase settings, its use in confirmatory and adaptive designs is rapidly expanding. Recent developments by Cetinyurek Yavuz et al., 2025 further explore the use of expert elicitation for hierarchical and multivariate priors, enabling richer modelling of treatment-effect uncertainty across multiple endpoints or populations.

3.8.4 Example

To illustrate how expert responses to elicitation questions are translated into a prior distribution and subsequently used to compute assurance, we consider a fictional Phase III oncology trial comparing a new treatment to standard of care. The primary endpoint is overall survival, and event times in the control arm are assumed to follow an exponential distribution with a median of 12 months, corresponding to a hazard rate of $\lambda_c = \log(2)/12$. The quantity of interest is the hazard ratio (HR) between the treatment and control arms, which is treated as uncertain. The trial is designed with 80% power to detect a hazard ratio of 0.75 at a one-sided significance level of $\alpha = 0.025$, requiring approximately 190 patients per arm.

3.8.4.1 Elicitation

A clinical expert with relevant experience is asked the following question:

> **Q.** *"Please provide the hazard ratio you would expect the treatment to achieve, corresponding to your 25th percentile, median, and 75th percentile of uncertainty."*

Suppose the expert provides quantiles of $\{0.65, 0.75, 0.88\}$ corresponding to the 25th percentile, median, and 75th percentile of their uncertainty about the hazard ratio.

These responses reflect the expert's belief that a hazard ratio of 0.75 is the most plausible value, with meaningful uncertainty in both directions. To translate these judgements into a prior distribution, a Gamma distribution is fitted by minimising the squared differences between the elicited and model-implied quantiles:

$$(a, b) = \arg\min_{\alpha,\beta>0} \sum_{i=1}^{3} \left[F_{\text{Gamma}(\alpha,\beta)}(q_i) - p_i\right]^2,$$

where $p_i \in \{0.25, 0.50, 0.75\}$ are the fixed quantile levels and $q_i \in \{0.65, 0.75, 0.88\}$ are the corresponding elicited values. Fitting yields HR $\sim$ Gamma$(19.98, 26.02)$, where the first and second parameters denote the shape and rate respectively.

The fitted distribution is shown in the left-hand panel of Figure 3.5, with the elicited quantiles and HR = 1 marked for reference — the latter illustrating the probability mass placed on the treatment being ineffective or harmful. The right-hand panel shows the CDF of the fitted distribution alongside the elicited quantile pairs, confirming the goodness of fit. The expert would then be asked to review the fitted distribution and revise their judgements if it does not adequately reflect their beliefs, following the iterative feedback approach described in Section 3.6.1.

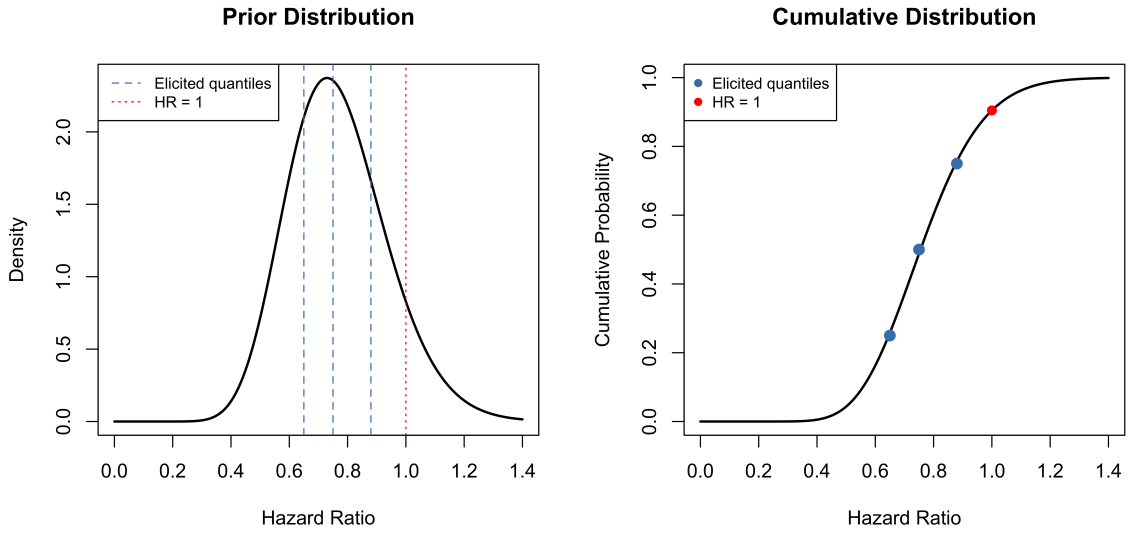


Figure 3.5: *Left: fitted Gamma(19.98, 26.02) prior for the hazard ratio, based on the elicited quantiles* $\{0.65, 0.75, 0.88\}$. *Vertical dashed lines indicate the 25th percentile, median, and 75th percentile of the fitted distribution; the dotted red line marks HR* $= 1$. *Right: CDF of the fitted prior with the elicited quantile pairs overlaid, confirming the goodness of fit.*

3.8.4.2 Assurance

To illustrate the full pipeline, we compute the assurance for this trial design by substituting the elicited prior into the assurance calculation framework described in Chapter 2. At each simulation replicate, a hazard ratio is sampled from the fitted Gamma prior, event times are generated for both arms under the exponential model, and the log-rank test is performed. The proportion of replicates in which the null hypothesis is rejected gives the assurance estimate. This yields an assurance of approximately 65%, notably lower than the nominal power of 80%. This reduction reflects the additional uncertainty about the true hazard ratio captured by the prior — the expert's judgement places meaningful probability on hazard ratios closer to 1, where the trial would be underpowered. This example demonstrates how elicitation and assurance together provide a more realistic assessment of the probability of trial success than a conventional power calculation alone.

3.8.5 Advantages and Considerations

The combination of elicitation and assurance offers several practical advantages:

- It allows decision-makers to explicitly incorporate expert knowledge about treatment plausibility and clinical relevance.
- It makes assumptions about treatment effects transparent and subject to scrutiny, facilitating regulatory discussion.
- It enables sensitivity analyses that explore how assurance varies under different prior scenarios, supporting more informed "go/no-go" decisions.

However, the approach also demands careful attention to transparency and documentation. Each step, from expert selection to prior fitting, should be recorded and justified, with clear rationale for the chosen elicitation framework and parameterisation. Elicited priors should be validated with the contributing experts and sensitivity analyses should be reported to assess the robustness of the assurance results.

3.8.6 Future Directions

Recent methodological work suggests several promising directions for the integration of elicitation and assurance. Advances in interactive software and visual tools (such as the SHELF R package) have simplified the elicitation process.

Moreover, hybrid approaches combining elicited priors with historical or real-world evidence are becoming increasingly common, offering a more balanced compromise between expert judgement and data-driven inference. Continued methodological refinement, along with empirical evaluation of how elicited priors perform in practice, will further enhance the credibility and uptake of assurance-based decision frameworks in industry.

In summary, elicitation provides a coherent and defensible means of specifying priors for assurance calculations, particularly when data are limited. When properly implemented, it ensures that Bayesian design evaluations reflect both statistical uncertainty and expert understanding, strengthening the link between quantitative design and clinical realism.

3.9 Summary

This chapter has focused on expert elicitation as a structured and defensible method for constructing prior distributions, which form a central component of assurance-based clinical trial design. We introduced the conceptual foundations of elicitation and its role in complementing empirical data, particularly in situations where evidence is sparse, uncertain, or heterogeneous.

Several established frameworks, such as SHELF, IDEA, Cooke's Classical Method, and the Delphi technique, were reviewed to illustrate how expert beliefs can be systematically captured, quantified, and aggregated. Methods for eliciting univariate and multivariate distributions were discussed, with emphasis on practical elicitation formats (such as quantiles and trial roulette), fitting procedures for translating expert inputs into formal probability distributions, and strategies for combining multiple expert judgements through mathematical or behavioural aggregation.

We also examined the cognitive and procedural challenges associated with elicitation, including biases such as anchoring and overconfidence, and highlighted best practices for expert training, facilitation, and transparency. The final section explored how elicitation methods can be directly integrated with assurance calculations, providing a coherent framework for quantifying trial success while fully accounting for both statistical and epistemic uncertainty.

By translating expert judgement into quantitative priors, elicitation links clinical insight directly to statistical design. Within Bayesian assurance frameworks, this yields a more realistic assessment of trial success that reflects both empirical evidence and domain expertise.

In the next chapter, we extend these ideas by applying assurance methods to more complex design settings, including trials where delayed treatment effects are anticipated. This builds on the principles established here, demonstrating how elicited priors can inform robust, transparent, and decision-relevant trial planning across a range of practical contexts.

Chapter 4

Assurance for Trials With a Delayed Treatment Effect[1]

4.1 Introduction

Immuno-oncology (IO) is a rapidly evolving field in anticancer drug development. In many IO trials, time-to-event outcomes such as overall survival (OS) and progression-free survival (PFS) exhibit nonproportional hazards (NPH), meaning that the relative treatment effect changes over time rather than remaining constant. This pattern violates the standard proportional hazards (PH) assumption that underpins conventional survival analyses.

A particularly important and clinically relevant form of NPH is a delayed treatment effect (DTE), where the survival curves for treatment and control remain similar for an initial period before separating later in follow-up. Such behaviour has been frequently observed in immunotherapy trials, for example, in CheckMate 017 (Brahmer, Reckamp, et al., 2015) and in several other anti–PD-1/PD-L1 studies summarised by Mukhopadhyay et al., 2022. In that review of 63 confirmatory randomised controlled trials, 15 exhibited evidence of nonproportional hazards, either through crossing survival curves or through an initial delay before separation (Rizvi et al., 2020; Herbst, Giaccone, et al., 2020; Shitara et al., 2020).

Designing trials that anticipate a DTE poses several challenges. The investigator must make assumptions about:

1. Whether a delay in treatment effect will occur;

[1]The research presented in this chapter was published in: Salsbury et al., 2024

2. The duration of the delay before benefit is observed; and
3. The magnitude of the eventual treatment benefit.

Incorrect assumptions about any of these aspects can substantially reduce statistical power, leading to underpowered studies that fail to detect meaningful treatment effects (Fine, 2007; Chen, 2013).

To address these challenges, we propose incorporating expert elicitation within the assurance framework to account for uncertainty in both the occurrence and duration of delayed effects. By formally integrating clinical knowledge into the design phase, this approach supports more realistic and transparent estimates of the probability of trial success.

4.2 Aims of the Chapter

This chapter develops a methodological and practical framework for calculating assurance in the design of survival trials that may exhibit delayed treatment effects. The specific aims are to:

1. Introduce the concept of delayed treatment effects and describe their implications for survival analysis in clinical trials;
2. Present parameterisations that appropriately model such effects within a time-to-event framework;
3. Identify the key parameters for which expert elicitation is required and describe how prior distributions can be constructed;
4. Demonstrate how assurance can be calculated under these models, explicitly accounting for uncertainty in the delay and effect size; and
5. Introduce and illustrate the accompanying open-source R package developed to perform these calculations.

4.3 Delayed Treatment Effects

In a typical survival trial, patients are randomised to receive either the experimental treatment or control. Let $h_c(t)$ and $h_e(t)$ denote the hazard functions for the control

and experimental treatment groups, respectively. Under the proportional hazards (PH) assumption, the treatment effect is constant over time such that:

$$h_e(t) = kh_c(t),$$

where k is a constant hazard ratio.

However, in many IO trials this assumption does not hold. Instead, patients in both groups may experience similar hazards during an initial period, followed by a reduction in the treatment arm once the immune response is activated. This pattern defines a delayed treatment effect (DTE).

We represent this behaviour using a piecewise hazard function:

$$h_e(t) = \begin{cases} h_c(t), & t \leq \tau, \\ h_e^*(t), & t > \tau, \end{cases} \tag{4.1}$$

where τ denotes the time of onset of treatment benefit and $h_e^*(t) < h_c(t)$ thereafter (Fine, 2007).

An example is shown in Figure 4.1, which presents a reconstructed Kaplan–Meier curve from the Phase III CheckMate 141 trial (Yen et al., 2020). The survival curves follow the same trajectory for approximately three months before separating, indicating the emergence of a treatment effect.

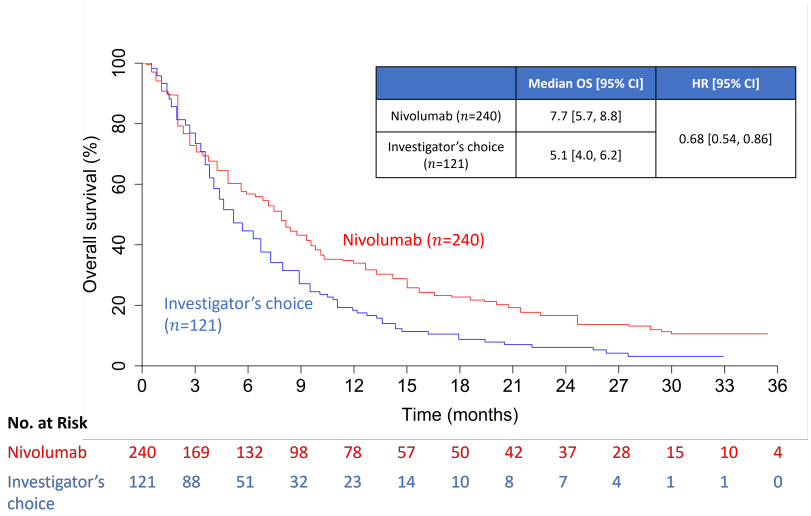


Figure 4.1: *Reconstructed Kaplan–Meier plot from the Phase III CheckMate 141 trial (Yen et al., 2020), demonstrating a delayed treatment effect (DTE). The control and treatment arms follow similar trajectories for the first three months before the curves diverge.*

Trials affected by delayed effects raise several design and analytical issues (Bardo et al., 2024). The primary estimand must remain clinically interpretable, the choice of hypothesis test must accommodate potential violations of the PH assumption (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2019; Jiménez et al., 2018), and the sample size or target event count must be robust to uncertainty about whether and when a delay will occur.

Various methods have been proposed to address these issues at the analysis stage. Weighted log-rank tests, for instance, place greater emphasis on later events when a treatment delay is expected (Fleming and Harrington, 1981). The restricted mean survival time (RMST) provides an alternative summary measure of treatment effect that does not rely on the proportional hazards assumption; it represents the expected survival time up to a pre-specified truncation time, calculated as the area under the survival curve (Royston and Parmar, 2013).

Despite these analytical advances, assurance-based methods for trial design under delayed treatment effects remain unexplored. Incorporating assurance offers a coherent way to quantify the overall probability of trial success while explicitly integrating uncertainty about both the delay and the eventual treatment effect. This chapter develops and illustrates such an approach.

4.4 Parameterisation for Delayed Treatment Effects

To calculate the assurance for a clinical trial with a delayed treatment effect (DTE), we must first define an appropriate parameterisation of the survival process. In survival trials, it is common to model event times in both the control and treatment arms using parametric distributions such as the exponential or Weibull distributions (Collett, 2015). However, when there is a delay in the onset of treatment effect, a more flexible model is required to capture the change in the hazard function over time.

We adopt a piecewise model, which allows for different hazard functions in different time intervals. For modelling purposes, we consider the exponential and Weibull distributions. Although more complex models, such as the log-normal, gamma, or log-logistic, could be used, these often add unnecessary complexity without substantially improving the model's ability to describe the anticipated behaviour of the trial. The exponential and Weibull distributions therefore offer an effective balance between flexibility and interpretability.

4.4.1 Exponential Model

We assume that event times in the control arm follow an exponential distribution with constant hazard rate λ_c. The corresponding survival function is given by:

$$S_c(t) = \exp(-\lambda_c t). \tag{4.2}$$

For the treatment arm, we model the delayed onset of the treatment effect using a piecewise exponential survival function. Prior to a fixed delay time τ, patients in the treatment group are assumed to experience the same hazard as those in the control group. After time τ, the treatment effect manifests as a change in the hazard rate from λ_c to λ_e. The survival function for the treatment group is thus defined as:

$$S_e(t) = \begin{cases} \exp(-\lambda_c t), & t \leq \tau, \\ \exp(-\lambda_c \tau - \lambda_e (t - \tau)), & t > \tau. \end{cases} \tag{4.3}$$

Under this parameterisation, the hazard ratio (HR) between the treatment and control arms is time-dependent and given by:

$$\mathrm{HR}(t) = \begin{cases} 1, & t \leq \tau, \\ \dfrac{\lambda_e}{\lambda_c}, & t > \tau. \end{cases} \tag{4.4}$$

This simple parameterisation allows the hazard to remain constant before the delay, and then decrease after τ, reflecting the emergence of the treatment benefit.

4.4.2 Weibull Model

To allow greater flexibility in the shape of the hazard function, we can alternatively assume that survival times in the control arm follow a Weibull distribution with survival function:

$$S_c(t) = \exp\{-(\lambda_c t)^{\gamma_c}\}, \tag{4.5}$$

where $\lambda_c > 0$ is a scale parameter and $\gamma_c > 0$ is a shape parameter.

Analogous to the exponential case, we model the delayed onset of treatment effect using a piecewise Weibull function. Prior to the delay time τ, patients in the treatment group have the same hazard as those in the control group. The treatment group survival

function is therefore defined as:

$$S_e(t) = \begin{cases} \exp\{-(\lambda_c t)^{\gamma_c}\}, & t \leq \tau, \\ \exp\{-(\lambda_c \tau)^{\gamma_c} - \lambda_e^{\gamma_e}(t^{\gamma_e} - \tau^{\gamma_e})\}, & t > \tau. \end{cases} \tag{4.6}$$

Under this parameterisation, the hazard ratio (HR) between the treatment and control arms is also time-dependent and given by:

$$\text{HR}(t) = \begin{cases} 1, & t \leq \tau, \\ \dfrac{\gamma_e \lambda_e^{\gamma_e} t^{\gamma_e - 1}}{\gamma_c \lambda_c^{\gamma_c} t^{\gamma_c - 1}}, & t > \tau. \end{cases} \tag{4.7}$$

This piecewise specification provides a flexible and interpretable way to model delayed treatment effects, forming the basis for constructing priors and calculating assurance in subsequent sections.

4.5 Constructing the Prior Distributions

From Sections 4.4.1 and 4.4.2, we see that there are three and five unknown parameters, respectively. In the exponential case, we have λ_c, λ_e, and τ. In the Weibull case, we have these three parameters, plus γ_c and γ_e. To calculate assurance, prior distributions are required for each of these parameters. In this section, we propose a practical method for constructing these priors.

It is important to note that although the questions below define the specific judgements required for the prior distributions, we would assume that the experts have already received some preparatory training, as described in Section 3.7.3, and have access to facilitator support. In addition, they are able to obtain clarification on modelling assumptions and receive structured feedback during the elicitation process.

4.5.1 Priors for the Control Group

In the exponential model, a single parameter λ_c governs survival in the control arm. In the Weibull model, two parameters are required: λ_c and γ_c. There are two main ways to obtain prior distributions for these control parameters.

4.5.1.1 Using historical information

If relevant historical data are available for the control intervention, these data can be used to inform a prior of the form

$$\pi(\boldsymbol{\theta_c} \mid \boldsymbol{x}_{\text{hist}}),$$

where $\boldsymbol{\theta_c}$ represents the control parameters and $\boldsymbol{x}_{\text{hist}}$ denotes the historical evidence. Section 2.7.1 describes general approaches for constructing such priors. Specifically for time-to-event data, Bertsche et al., 2019 and Roychoudhury and Neuenschwander, 2020 extend the method of Schmidli et al., 2014 to derive priors for control-arm survival functions.

One widely used approach is the Meta-Analytic-Predictive (MAP) prior (Neuenschwander et al., 2010; Schmidli et al., 2014). Given H historical studies, the MAP prior is constructed by modelling study-specific control parameters as exchangeable across studies and the future trial. In both cases, the MAP prior for the future control parameters is the posterior predictive distribution obtained by marginalising over the historical data and hyperparameters. Weakly informative priors are assigned to $\boldsymbol{\mu}$ and the variance components, with sensitivity analyses recommended when the number of historical studies is small.

4.5.1.1.1 Exponential Model

For the exponential model, the log hazard rates are modelled as exchangeable:

$$\log(\lambda_1), \ldots, \log(\lambda_H), \log(\lambda_C) \sim \mathcal{N}(\mu, \omega^2),$$

where μ represents the overall mean log hazard rate and ω captures between-study variability.

4.5.1.1.2 Weibull Model

For the Weibull model, the log-transformed parameter vectors $\boldsymbol{\phi}_h = (\log(\lambda_h), \log(\gamma_h))^\top$ are modelled jointly as exchangeable:

$$\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_H, \boldsymbol{\phi}_C \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ represents the overall mean vector and $\boldsymbol{\Sigma}$ captures between-study variability in both parameters and their correlation.

4.5.1.2 Using expert elicitation

When historical data are sparse, heterogeneous, or of limited relevance, prior distributions may instead be constructed from elicited expert beliefs. Methods for eliciting

priors for common survival models have been proposed in the literature; for example, Ren and Oakley, 2014 provide structured approaches for obtaining expert judgements for both exponential and Weibull models.

Building on this work, we outline below a specific set of elicitation questions required to construct priors for the exponential and Weibull survival models used in this thesis. In the elicitation questions below, the quantile levels are fixed in advance by the facilitator. Here, we choose to elicit the 25th percentile, median, and 75th percentile, though in practice the choice of quantiles is at the discretion of the facilitator and should reflect the needs of the elicitation and the expertise of the participants.

4.5.1.2.1 Model Choice

Before proceeding with model-specific elicitation, it is useful to establish which survival model is most appropriate for the control arm. To this end, we first ask:

> **M1.** *"Do you expect the hazard of the event in the control group to remain approximately constant over time, or do you expect it to increase or decrease?"*

If the expert believes the hazard is approximately constant, an exponential model is appropriate and the elicitation proceeds as described in Section 4.5.1.2.2. If the expert believes the hazard changes over time — for example, increasing as the disease progresses or decreasing as early non-responders are removed from the risk set — a Weibull model is more appropriate, and the elicitation proceeds as described in Section 4.5.1.2.3. In cases where the expert is uncertain, both models can be considered and the sensitivity of the assurance to the choice of model explored.

4.5.1.2.2 Exponential Model

For the exponential case, we need to elicit beliefs about λ_c. We do so by considering a clinically relevant landmark time t_1 and the corresponding survival probability $s_1 = S(t_1)$. The quantity s_1 is modelled using a Beta(a, b) prior distribution. To elicit the values for the hyperparameters a and b, we ask:

> **E1.** *"Consider a clinically relevant landmark time t_1. Please provide the survival probability at time t_1 corresponding to your 25th percentile, median, and 75th percentile of uncertainty."*

Let the elicited quantile pairs be denoted by $\{(p_i, q_i)\}_{i=1}^{3}$, where $p_i \in \{0.25, 0.50, 0.75\}$ represents the fixed quantile level and $q_i \in (0, 1)$ the corresponding survival probability

provided in response to **E1**. The hyperparameters (a, b) of the Beta distribution for s_1 are then obtained by minimising the squared differences between the elicited and model-implied quantiles:

$$(a, b) = \arg\min_{\alpha,\beta>0} \sum_{i=1}^{3} \left[F_{\text{Beta}(\alpha,\beta)}(q_i) - p_i\right]^2,$$

where $F_{\text{Beta}(\alpha,\beta)}$ denotes the cumulative distribution function of a Beta(α, β) distribution. Once a value of s_1 is sampled from the fitted Beta(a, b) distribution, the exponential rate parameter for the control group is calculated as

$$\lambda_c = -\frac{\log(s_1)}{t_1}.$$

4.5.1.2.3 Weibull Model

For the Weibull case, we need to elicit beliefs about λ_c and γ_c. We do so by considering two clinically relevant landmark times, t_1, t_2 (with $t_1 < t_2$) and the corresponding survival probabilities $s_1 = S(t_1)$ and $s_2 = S(t_2)$.

To identify both parameters, we elicit uncertainty about s_1 and the decrease in survival between t_1 and t_2, defined as $\delta = s_1 - s_2$. The quantities s_1 and δ are modelled using independent Beta priors:

$$s_1 \sim \text{Beta}(a, b), \quad \delta \sim \text{Beta}(c, d).$$

To elicit the corresponding hyperparameters, we ask:

> **W1.** *"Consider a clinically relevant time point t_1. Please provide the survival probability at time t_1 corresponding to your 25th percentile, median, and 75th percentile of uncertainty."*

> **W2.** *"Consider a second clinically relevant time point $t_2 > t_1$. Please provide the decrease in survival between t_1 and t_2 corresponding to your 25th percentile, median, and 75th percentile of uncertainty."*

Let the elicited quantile pairs in response to **W1** be denoted by $\{(p_i, q_i)\}_{i=1}^3$, where $p_i \in \{0.25, 0.50, 0.75\}$ represents the fixed quantile level and $q_i \in (0, 1)$ the corresponding survival probability at t_1. Let the elicited quantile pairs in response to **W2** be denoted by $\{(p_j, r_j)\}_{j=1}^3$, where $p_j \in \{0.25, 0.50, 0.75\}$ and $r_j \in (0, 1)$ the corresponding elicited

decrease in survival between t_1 and t_2. The Beta hyperparameters are obtained by minimising the squared differences between the elicited and model-implied quantiles:

$$(a, b) = \arg\min_{\alpha,\beta>0} \sum_{i=1}^{3} \left[F_{\text{Beta}(\alpha,\beta)}(q_i) - p_i\right]^2, \qquad (c, d) = \arg\min_{\alpha,\beta>0} \sum_{j=1}^{3} \left[F_{\text{Beta}(\alpha,\beta)}(r_j) - p_j\right]^2.$$

Once s_1 and δ are sampled from their respective fitted Beta distributions, the Weibull parameters for the control group are then computed as

$$\lambda_c = \frac{[-\log(s_1)]^{1/\gamma_c}}{t_1}, \qquad \gamma_c = \frac{\log\left[\frac{\log(s_1)}{\log(s_1-\delta)}\right]}{\log(t_1/t_2)}.$$

4.5.2 Priors for the Treatment Group

We assume that expert beliefs have already been elicited for the parameters governing the control group, $\boldsymbol{\theta}_C$. We now turn to beliefs about the treatment-group parameters, $\boldsymbol{\theta}_E$.

4.5.2.1 Prior for the Length of Delay

We begin by eliciting the delay time, τ. The presence of a delayed treatment effect (DTE) presupposes that the treatment has some beneficial impact, although experts may be uncertain about whether such an effect exists. Accordingly, we elicit both (i) the probability that any treatment effect occurs, and (ii) the distribution of the delay time τ conditional on such an effect.

To ensure clarity, we define a treatment effect as any separation between the population-level survival curves for the treatment and control groups. Let $\mathcal{S}$ denote the proposition that such a separation exists, referring specifically to the underlying survival functions rather than sample-based Kaplan–Meier estimates. The expert is asked to provide their probability that $\mathcal{S}$ is true, denoted

$$P_{\mathcal{S}} \in [0, 1].$$

The first elicitation question is therefore:

> **Q1:** *"What is your probability that the treatment has any impact on survival, that is, that the population survival curves for the treatment and control groups separate at some point in time?"*

Given $\mathcal{S}$, we allow for the possibility that no delay is present. We specify the prior for τ as

$$\tau \mid \mathcal{S} = \begin{cases} 0, & \text{with probability } 1 - P_{\text{DTE}}, \\ D_{\text{delay}}, & \text{with probability } P_{\text{DTE}}, \end{cases} \tag{4.8}$$

where $D_{\text{delay}} \sim \text{Gamma}(a, b)$. Although any non-negative distribution could be used for D_{delay}, the Gamma distribution is assumed to offer sufficient flexibility for plausible delay times.

To elicit the parameters of this prior, we ask:

> **Q2:** *"Suppose that the population survival curves do separate, what is your probability that there is a delay before they separate?"*

Let $P_{\text{DTE}} \in [0, 1]$ denote the expert's answer to **Q2**.

Conditional on a delay being present, we next ask:

> **Q3:** *"Suppose the population survival curves separate with a delay. Please provide the delay time corresponding to your 25th percentile, median, and 75th percentile of uncertainty about the duration of this delay."*

Let the elicited quantile–value pairs be $\{(p_i, q_i)\}_{i=1}^{3}$, where $p_i \in \{0.25, 0.50, 0.75\}$ is the fixed quantile level and $q_i > 0$ the corresponding delay time.

The hyperparameters (a, b) of the Gamma distribution for D_{delay} are obtained by minimising the squared deviation between the elicited and model-implied quantiles:

$$(a, b) = \arg\min_{\alpha,\beta>0} \sum_{i=1}^{n} \left[F_{\Gamma(\alpha,\beta)}(q_i) - p_i \right]^2,$$

where $F_{\Gamma(\alpha,\beta)}$ denotes the CDF of a Gamma(α, β) distribution.

4.5.2.2 Priors for the Treatment Effect

Having elicited beliefs about the length of delay, τ, we now turn to eliciting beliefs about the treatment effect. In the exponential case, the treatment group requires a prior for λ_e; in the Weibull case, for λ_e and γ_e. Rather than eliciting beliefs directly about these parameters, which are often unintuitive, we elicit beliefs about observable quantities (Kadane and Wolfson, 1998), such as:

- Median survival time on the experimental treatment;
- Survival probability at a clinically relevant time t;
- The magnitude and timing of the maximum separation between survival curves.

If experts are more comfortable reasoning in terms of hazard ratios, questions may instead be framed as:

- "What is your best estimate of the hazard ratio at time τ?"
- "What is the lowest hazard ratio you would expect, and when might it occur?"

In practice, we focus on eliciting beliefs about the hazard ratio once the treatment effect begins to act. We term this quantity the *post-delay hazard ratio*, denoted HR^*.

For the Weibull model, two parameters (λ_e, γ_e) must be elicited, which can be burdensome for experts. To simplify, we assume $\gamma_e = \gamma_c$, a simplification explored in Section 4.8. Under this assumption, the hazard ratio simplifies to a piecewise-constant form:

$$\mathrm{HR}(t) = \begin{cases} 1, & t \leq \tau, \\ \left(\dfrac{\lambda_e}{\lambda_c}\right)^{\gamma_c}, & t > \tau, \end{cases} \tag{4.9}$$

and rearranging for λ_e gives:

$$\lambda_e = \lambda_c \, \mathrm{HR}^{1/\gamma_c}. \tag{4.10}$$

Conditional on λ_c and γ_c, we elicit a distribution for the post-delay hazard ratio HR^*, from which a prior for λ_e can be derived. We assume that the treatment effect, represented by HR^*, is conditionally independent of the control group parameters λ_c and γ_c.

We propose the following prior for the post-delay hazard ratio:

$$\mathrm{HR}^* \sim \mathrm{Gamma}(c, d), \tag{4.11}$$

although other non-negative distributions could also be considered.

To elicit the parameters c and d, we ask:

> **Q4:** *"Suppose that the population survival curves separate. Please provide the hazard ratio, once the experimental treatment begins to take effect, corresponding to your 25th percentile, median, and 75th percentile of uncertainty."*

Let the elicited quantile pairs be denoted by $\{(p_i, q_i)\}_{i=1}^{3}$, where $p_i \in \{0.25, 0.50, 0.75\}$ represents the fixed quantile level and $q_i > 0$ the corresponding value of HR^* provided in response to **Q4**. The hyperparameters (c, d) are then estimated by minimising the squared difference between elicited and model-implied quantiles:

$$(c, d) = \arg\min_{\alpha,\beta>0} \sum_{i=1}^{3} \left[F_{\mathrm{Gamma}(\alpha,\beta)}(q_i) - p_i\right]^2,$$

where $F_{\mathrm{Gamma}(\alpha,\beta)}$ denotes the cumulative distribution function of a $\mathrm{Gamma}(\alpha, \beta)$ distribution.

4.5.3 Full Parameterisation

Conditional on the existence of a treatment effect ($\mathcal{S}$) and on any historical information about the control arm ($\boldsymbol{x}_{\mathrm{hist}}$), the joint prior distribution for all model parameters can be written as

$$\pi(\tau, \mathrm{HR}^*, \boldsymbol{\theta_c} \mid \mathcal{S}, \boldsymbol{x}_{\mathrm{hist}}) = \pi(\boldsymbol{\theta_c} \mid \boldsymbol{x}_{\mathrm{hist}})\, \pi(\tau \mid \mathcal{S})\, \pi(\mathrm{HR}^* \mid \mathcal{S}), \tag{4.12}$$

where $\boldsymbol{\theta_c} = (\lambda_c)$ under the exponential model and $\boldsymbol{\theta_c} = (\lambda_c, \gamma_c)$ under the Weibull model. This factorisation assumes that, given $\mathcal{S}$, the delay time τ and the post-delay hazard ratio HR^* are conditionally independent of the control-arm parameters.

If no treatment effect exists (i.e. $\mathcal{S}$ is false), the treatment and control survival curves coincide. For computational convenience, we enforce this by fixing

$$\tau = 0, \qquad \mathrm{HR}^* = 1.$$

In this setting, the prior for the control-group parameters is unaffected by the assumption about $\mathcal{S}$, and therefore

$$\pi(\boldsymbol{\theta_c} \mid \mathcal{S}, \boldsymbol{x}_{\mathrm{hist}}) = \pi(\boldsymbol{\theta_c} \mid \boldsymbol{x}_{\mathrm{hist}}).$$

Throughout, we assume conditional independence between HR^* and $\tau \mid \mathcal{S}$. This simplifying assumption eases elicitation and computation, although dependence could be introduced if experts believe, for instance, that larger treatment effects tend to occur after longer delays. Section 3.4 describes how such dependence structures may be incorporated if required.

The complete elicitation procedure is summarised in Figure 4.2.

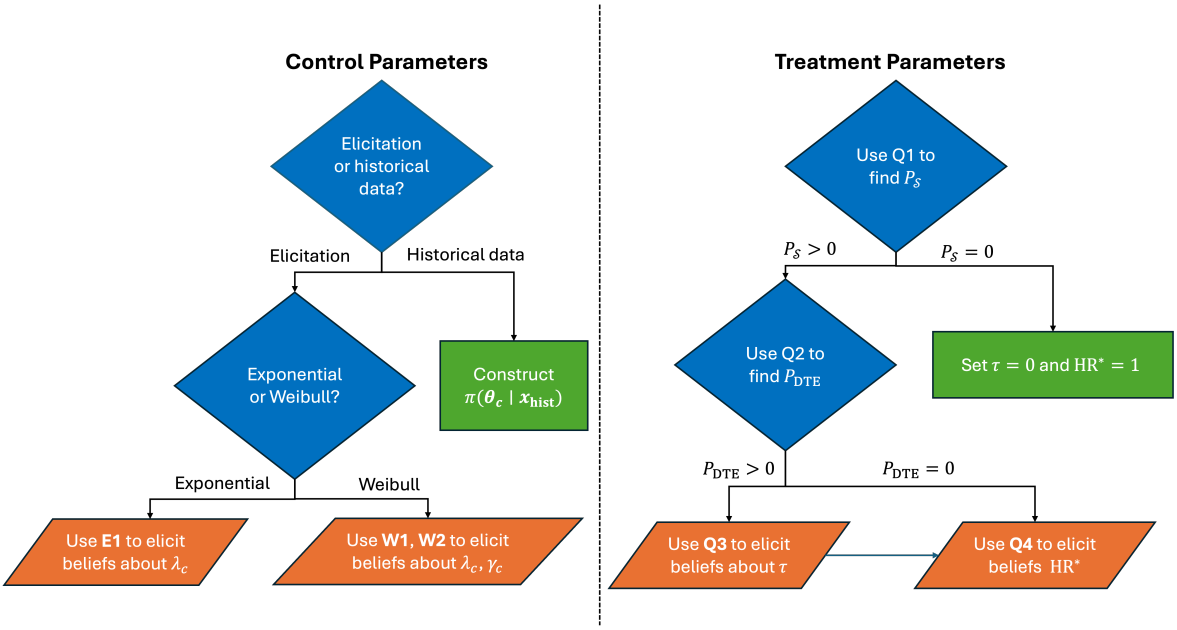


Figure 4.2: *The proposed elicitation scheme, as described in Sections 4.5.1 and 4.5.2.*

4.6 Calculating Assurance

The elicited prior distributions described in the previous sections are now used to calculate the assurance for a range of possible trial designs. The computational framework for this calculation is outlined in Algorithm 5.

Algorithm 5 Assurance calculation under a delayed treatment effect (DTE)

Inputs: Per-arm sample sizes n_c, n_e; priors $\pi(\boldsymbol{\theta_c})$, $\pi(\tau)$, $\pi(\text{HR}^*)$; separation probability $P_{\mathcal{S}}$; delayed-effect probability P_{DTE}; recruitment model R; censoring model C; primary analysis A; number of replicates N
for $i = 1, \ldots, N$ **do**
  Sample $\boldsymbol{\theta_{c,i}} \sim \pi(\boldsymbol{\theta_c})$
  if $u \sim \text{Uniform}(0,1) \geq P_{\mathcal{S}}$ **then**
    Set $\tau_i = 0$ and $\text{HR}_i^* = 1$ ▷ No treatment effect
  else if $u \sim \text{Uniform}(0,1) < P_{\text{DTE}}$ **then**
    Sample $\tau_i \sim \pi(\tau)$ and $\text{HR}_i^* \sim \pi(\text{HR}^*)$ ▷ Delayed treatment effect
  else
    Set $\tau_i = 0$ and sample $\text{HR}_i^* \sim \pi(\text{HR}^*)$ ▷ Immediate treatment effect
  end if
  Generate control event times using $\boldsymbol{\theta_{c,i}}$
  Generate treatment event times using $\boldsymbol{\theta_{c,i}}$, τ_i, HR_i^*
  Apply recruitment model R and censoring model C
  Perform primary analysis A and set $U_i = \mathbf{1}$ (if analysis successful)
end for
Estimate assurance, where S denotes the event of a successful analysis:

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

Several design parameters influence the calculated assurance, including the sample sizes in the control (n_c) and treatment (n_e) arms, and the total number of events E. These can be varied to reflect the operational and logistical constraints associated with running a clinical trial. Each parameter impacts the assurance differently. For example, for fixed n_c and ne, increasing the target number of events E will almost always increase assurance, but at the cost of a longer trial duration. It is therefore important to explore alternative design configurations to identify an optimal balance between statistical robustness and feasibility.

Censoring arises naturally in survival trials and must be accounted for in the simulation. Administrative censoring occurs when the trial ends before all patients have experienced the event, and is determined by the recruitment model and the planned analysis time. Loss to follow-up, whereby patients withdraw or become unreachable before the trial concludes, can also be incorporated by specifying a dropout model within the censoring component (C) of the algorithm. The impact of censoring on assurance is captured implicitly through its effect on the observed number of events, which in turn determines the power of the log-rank test.

The general algorithmic framework is flexible and can be adapted to the specific needs of a proposed clinical trial. Although a log-rank test is the default method of analysis, other options such as a weighted log-rank test or a comparison based on restricted mean survival time (RMST) can be substituted. This flexibility ensures that the assurance calculation accurately reflects the planned analysis method in the trial protocol. Similarly, more complex design elements can be incorporated, such as non-uniform recruitment schedules (e.g., ramp-up or staggered enrolment) and interim analyses (discussed in detail in Chapter 6). The simulation-based structure of the algorithm means that such extensions require minimal modification to the core procedure.

Finally, increasing the number of replicates N reduces Monte Carlo variability in the assurance estimate, producing a more stable result. The Monte Carlo uncertainty in the estimate can be quantified using the 95% Monte Carlo interval reported alongside the assurance value.

4.6.1 Calculating Assurance using the `DTEAssurance` package

Algorithm 5 is implemented to calculate assurance using the `DTEAssurance` R package, with the function `DTEAssurance::calc_dte_assurance()`. The function requires several user-specified inputs:

- `n_c`: Vector of control group sample sizes;
- `n_e`: Vector of treatment group sample sizes;
- `control_model`: List specifying information about the control arm survival distribution;
- `effect_model`: List defining beliefs about the treatment effect;
- `censoring_model`: List describing the censoring mechanism;
- `recruitment_model`: List specifying the recruitment process;
- `analysis_model`: List describing the planned statistical test and decision rule;
- `n_sims`: Number of simulation iterations to perform.

The output is a list (of length `n_c`) containing:

- `assurance`: Estimated assurance (probability of success under prior uncertainty);

- `CI`: 95% Monte Carlo interval for the assurance estimate, reflecting simulation uncertainty;
- `duration`: Mean trial duration across simulations;
- `sample_size`: Mean sample size across simulations.

4.7 Illustrative Examples and R Implementation

This section illustrates the proposed methodology through two case studies. The first example employs an exponential survival model, while the second uses a Weibull model. Both examples demonstrate how elicited priors can be incorporated into assurance calculations and how the `DTEAssurance` R package supports this process.

4.7.1 Exponential Example

We first consider the design of a two-arm Phase III superiority trial evaluating the efficacy of a novel immuno-oncology (IO) drug compared with the standard of care, docetaxel, in patients with advanced non–small-cell lung cancer (NSCLC). Given the mechanism of IO treatments, a delayed treatment effect (DTE) is anticipated. The primary efficacy endpoint is overall survival (OS).

We assume uniform patient recruitment over 12 months, with 1:1 randomisation between arms. The data will be analysed using a one-sided log-rank test of size 2.5%. The final analysis will be conducted once 80% of the planned sample size has experienced the event, after which any remaining patients will be censored.

4.7.1.1 Prior for the Control Group

Historical data are available for docetaxel, the control-arm treatment. We use this prior information to construct a prior distribution for control-group survival (as introduced in Section 4.5.1.1.1). Following Bertsche et al., 2019, three clinical trials were identified where docetaxel served as the control arm: INTEREST (Kim, Hirsh, et al., 2008), ZODIAC (Herbst, Sun, et al., 2010), and REVEL (Garon et al., 2014). Individual patient data were reconstructed from the published Kaplan–Meier curves of these trials using the method proposed by Liu et al., 2021.

Figure 4.3 presents the reconstructed survival curves for each trial. We apply the Meta-Analytic Predictive (MAP) prior framework of Schmidli et al., 2014, as in Bertsche et

al., 2019, though in our case for OS rather than progression-free survival (see Appendix A.2 for further details).

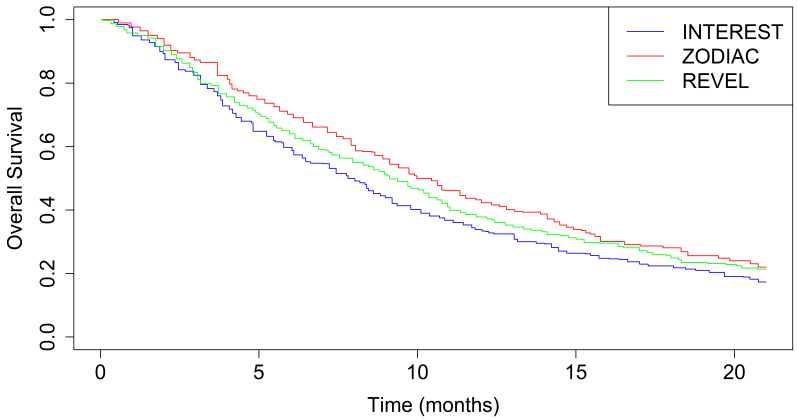


Figure 4.3: *Reconstructed Kaplan–Meier curves for the docetaxel arm in three trials: ZODIAC (Herbst, Sun, et al., 2010), REVEL (Garon et al., 2014), and INTEREST (Kim, Hirsh, et al., 2008). The similarity among the curves supports the assumption of exchangeability.*

Table 4.1 summarises the estimated exponential rates ($\hat{\lambda}_h$) for each trial, derived from the reconstructed data. Using these, we generate MCMC samples for the MAP prior for λ_c, shown in Figure 4.4a. Figure 4.4b overlays the three historical Kaplan–Meier curves with the median survival curve (solid black) and 90% credible interval (dashed black) corresponding to the MAP prior.

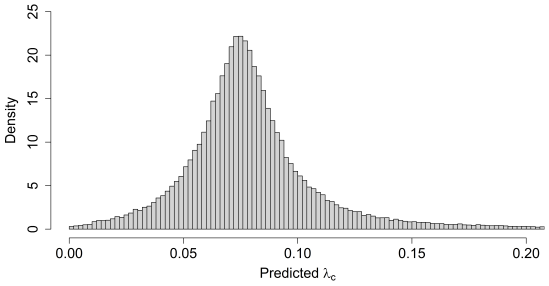
Study	Number of events (d_h)	Median OS (months)	$\hat{\lambda}_h$
INTEREST	574	7.81	0.089
ZODIAC	399	10.4	0.067
REVEL	451	9.12	0.076

Table 4.1: *Summary statistics for overall survival in the docetaxel arms of three published trials.*

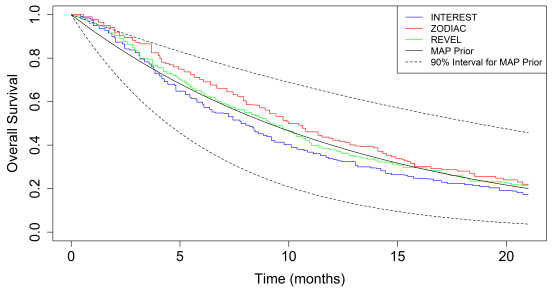
4.7.1.2 Prior for the Treatment Group

We now illustrate the elicitation of prior distributions for the treatment-group parameters, following the hierarchical structure shown in Figure 4.2.

For **Q1**, suppose the expert provides $P_S = 0.9$, indicating a high degree of belief that the population survival curves will eventually separate.



(a) *Histogram of MCMC samples from the MAP prior distribution for λ_c.*



(b) *Historical Kaplan–Meier curves (coloured) with the MAP prior median survival curve (solid black) and 90% credible interval (dashed black).*

Figure 4.4: *MAP prior distribution for λ_c (left) and corresponding survival curve implications (right).*

For **Q2**, suppose the expert specifies $P_{\text{DTE}} = 0.8$, reflecting confidence that, if a treatment effect exists, it will manifest only after a delay.

For **Q3**, suppose the expert provides the quantiles $\{0.25, 0.5, 0.75\}$ with corresponding delay times $\{3, 4, 5\}$ months. Fitting a Gamma distribution to these judgements yields

$$\tau \mid S \sim \begin{cases} 0, & \text{with probability } 0.2, \\ \text{Gamma}(7.29,\ 1.76), & \text{with probability } 0.8, \end{cases}$$

with model-implied quartiles 3.03, 3.95, and 5.05 months. These fitted values would typically be shown to the expert for validation and potential revision.

Figure 4.5 illustrates this fitting procedure as displayed in the Shiny application launched by `DTEAssurance::assurance_shiny_app()`.

Finally, for **Q4**, suppose the expert provides quantiles $\{0.25, 0.5, 0.75\}$ with corresponding values for the post-delay hazard ratio $\{0.55, 0.60, 0.70\}$. Fitting a Gamma distribution to these values gives

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

with fitted quartiles 0.54, 0.61, and 0.69.

4.7.1.3 Calculating Assurance

The elicited priors are used to calculate assurance via Algorithm 5. Figure 4.6 presents the resulting assurance curve together with two power curves. Both power curves use

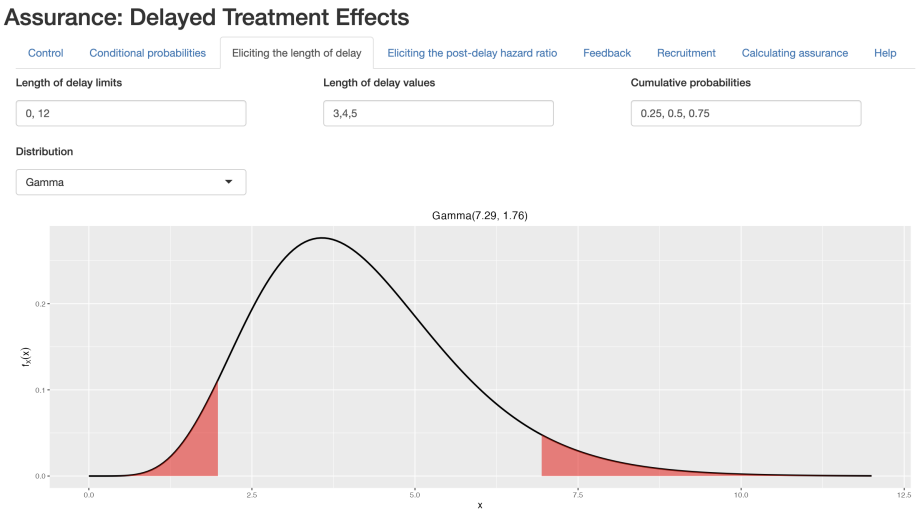


Figure 4.5: *Screenshot from the Shiny application launched by* `DTEAssurance::assurance_shiny_app()`*, illustrating the quantile-based elicitation and Gamma fitting of the delay time* τ*.*

point estimates for parameters ($\lambda_c = 0.077$, $\tau = 4$, $\text{HR}^* = 0.6$); one assumes no delay ($\tau = 0$). The corresponding parameter specifications are summarised in Table 4.2. Recruitment, analysis, and censoring models are held constant across all scenarios.

Method	λ_c	τ	HR^*	P_S	P_{DTE}
Assurance	MCMC sample	Ga(7.29, 1.76)	Ga(29.6, 47.8)	0.9	0.8
Power	0.077	4	0.6	1	1
Power (no delay)	0.077	0	0.6	1	0

Table 4.2: *Parameter specifications for the three scenarios presented in Figure 4.6.*

Both power calculations are notably more optimistic than the assurance estimate, demonstrating the importance of incorporating uncertainty into trial design. Assurance should not be used solely to set sample size, but as part of a broader decision framework that considers trial feasibility, expected duration, number of events, and operating characteristics. Exploring alternative designs through assurance curves enables evidence-based decision-making in the planning phase.

4.7.2 Weibull Example

The second case study concerns a two-arm Phase III superiority trial comparing a novel IO therapy with the current standard of care, dacarbazine, in patients with

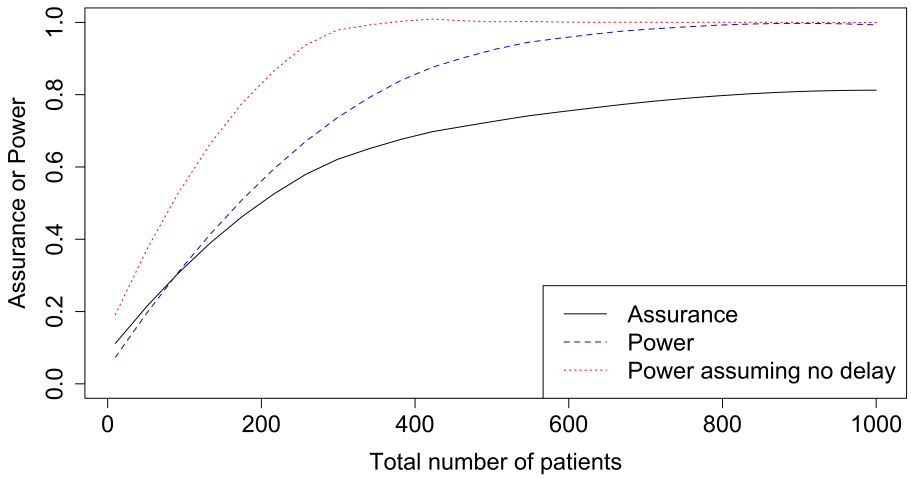


Figure 4.6: *Power and assurance curves for the exponential example (Section 4.7.1). The differences highlight the impact of parameter uncertainty on trial planning.*

advanced melanoma. As before, a DTE is anticipated and the primary endpoint is OS. Uniform recruitment is assumed over six months, with 2:1 randomisation in favour of the experimental arm. The primary analysis will be conducted using a log-rank test, once 70% of participants have experienced the event.

4.7.2.1 Prior for the Control Group

We assume the expert has prior experience with dacarbazine and is familiar with the relevant literature (Robert, Thomas, et al., 2011; Robert, Long, Brady, Dutriaux, Maio, et al., 2015; Ribas et al., 2013). The elicitation focuses on two quantities: the six-month survival probability $S_1(6)$ and the survival drop between six and twelve months,

$$\delta = S_1(6) - S_1(12).$$

As introduced in Section 4.5.1.2.3, we assume $S_1(6) \sim \text{Beta}(a, b)$ and $\delta \sim \text{Beta}(c, d)$.

Using **W1** and **W2**, we determine the hyperparameters a, b, c and d. Suppose the expert initially provides the quantiles $\{0.25, 0.5, 0.75\}$ with corresponding values $\{0.6, 0.7, 0.8\}$ for $S_1(6)$, and quantiles $\{0.25, 0.5, 0.75\}$ with values $\{0.2, 0.3, 0.4\}$ for δ.

Fitting Beta distributions to these judgements yields:

$$S_1(6) \sim \text{Beta}(6.64,\ 2.98), \qquad \delta \sim \text{Beta}(2.98,\ 6.64).$$

The implied prior is visualised in Figure 4.7, which shows the output produced by the Shiny application launched via `DTEAssurance::assurance_shiny_app()`. A large number of Weibull survival curves are sampled from the prior and the pointwise median, 2.5th, and 97.5th percentiles are displayed, providing a visual summary of the uncertainty in the control arm survival function.

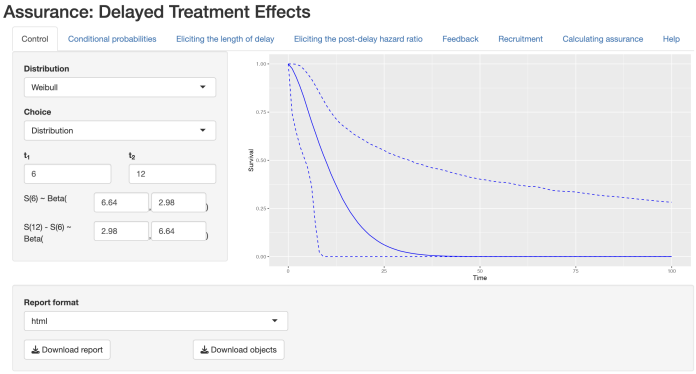


Figure 4.7: *Implied prior for the control-group survival parameters, as displayed in the Shiny application launched by* `DTEAssurance::assurance_shiny_app()`. *The plot shows the median survival curve (solid blue) with pointwise 2.5th and 97.5th percentiles (dashed blue).*

Suppose the expert considers the resulting pointwise intervals too wide. After reviewing the feedback provided by the application, they revise their judgements to reflect greater certainty. They retain the same quantiles but update the corresponding values: for $S_1(6)$, the values remain $\{0.6, 0.7, 0.8\}$, while for δ they are narrowed to $\{0.25, 0.3, 0.35\}$.

Fitting updated Beta distributions gives:

$$S_1(6) \sim \text{Beta}(26.7,\ 11.6), \qquad \delta \sim \text{Beta}(11.6,\ 26.7),$$

as shown in Figure 4.8.

4.7.2.2 Prior for the Treatment Group

Following the hierarchical elicitation framework in Figure 4.2, we now elicit the treatment-effect parameters using the same sequence of questions **Q1**–**Q4** introduced earlier.

Q1: The expert is first asked for their probability that the population survival curves for the treatment and control groups separate. They respond with $P_S = 1$, indicating certainty that a treatment effect exists.

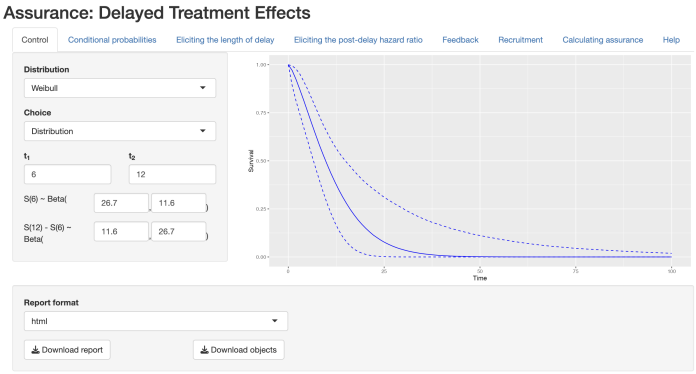


Figure 4.8: *Updated elicited prior for the control-group survival distribution, obtained after the expert revised their judgements to express reduced uncertainty. The resulting median survival curve (solid blue) displays correspondingly tighter pointwise 2.5th and 97.5th percentiles* (dashed blue).

Q2: Conditional on $\mathcal{S}$, the expert is asked for their probability that the effect exhibits a delay before separation occurs. They specify $P_{\text{DTE}} = 0.5$.

Q3: Conditional on a delay being present, the expert provides beliefs about the delay length. They specify a median of 3 months, with lower and upper quartiles of 1 and 4 months, respectively. Fitting a Gamma distribution to these elicited quantiles gives

$$D_{\text{delay}} \sim \text{Gamma}(1.14,\, 0.336),$$

so that the prior for the delay time is

$$\tau \mid \mathcal{S} = \begin{cases} 0, & \text{with probability } 0.5, \\ \text{Gamma}(1.14,\, 0.336), & \text{with probability } 0.5. \end{cases}$$

The model-implied quartiles (1.11, 2.47, 4.70) are then returned to the expert for confirmation.

Q4: Finally, the expert is asked for their beliefs about the post-delay hazard ratio HR^*. They provide a median of 0.5 and quartiles 0.4 and 0.6. Fitting a Gamma distribution to these judgements yields

$$\text{HR}^* \sim \text{Gamma}(11.4,\, 22.3),$$

which closely reproduces the elicited quantiles.

4.7.2.3 Calculating Assurance

Finally, Algorithm 5 is used to compute assurance across a range of sample sizes. The resulting assurance curve is shown in Figure 4.9.

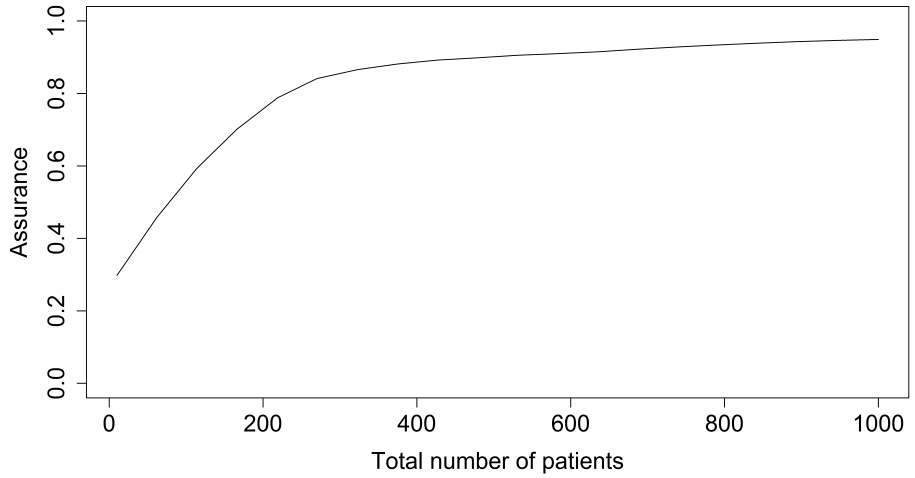


Figure 4.9: *Assurance curve for the Weibull example (Section 4.7.2).*

4.8 Simplified prior distribution: discussion

In this section, we assess the simplification made to the Weibull parameterisation by examining its impact on fitted survival curves and on design characteristics such as power and assurance. We also present an alternative, more flexible approach to calculating assurance, suitable for situations where the simplification may not be appropriate. During the elicitation, it is worth asking experts whether they believe the hazard ratio between the treatment and control groups will remain approximately constant after the treatment begins to take effect. If the expert has concerns, the flexible approach described in this section may be more suitable.

4.8.1 Robustness of the parameterisation

To make the elicitation process manageable for experts, we simplified the Weibull model by assuming $\gamma_e = \gamma_c$. This allowed us to focus on eliciting beliefs about the post-delay hazard ratio rather than requiring direct judgements about both the treatment and control shape parameters, a task that would have been more difficult and less intuitive.

To assess the implications of this simplification, we compared two parameterisations, denoted **Method A** and **Method B**. Both share the same control-arm Weibull model but differ in the treatment-arm formulation.

For the control group,

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

For the treatment group, Method A (the simplified form) is:

$$S_e(t) = \begin{cases} \exp\{-(\lambda_c t)^{\gamma_c}\}, & t \le \tau, \\ \exp\{-(\lambda_c \tau)^{\gamma_c} - \lambda_e^{\gamma_c}(t^{\gamma_c} - \tau^{\gamma_c})\}, & t > \tau, \end{cases}$$

and Method B (the full Weibull parameterisation) is:

$$S_e(t) = \begin{cases} \exp\{-(\lambda_c t)^{\gamma_c}\}, & t \le \tau, \\ \exp\{-(\lambda_c \tau)^{\gamma_c} - \lambda_e^{\gamma_e}(t^{\gamma_e} - \tau^{\gamma_e})\}, & t > \tau. \end{cases}$$

Hence, Method A includes four parameters $(\lambda_c, \lambda_e, \gamma_c, \tau)$, while Method B has five $(\lambda_c, \lambda_e, \gamma_c, \gamma_e, \tau)$.

To evaluate robustness, we analysed three trials in which delayed treatment effects were observed: CheckMate 017 (Brahmer, Reckamp, et al., 2015), CheckMate 141 (Yen et al., 2020), and the pooled analysis of CheckMate 017 and 057 (Borghaei, Gettinger, et al., 2021). Figure 4.10 displays the reconstructed Kaplan–Meier curves for these datasets.

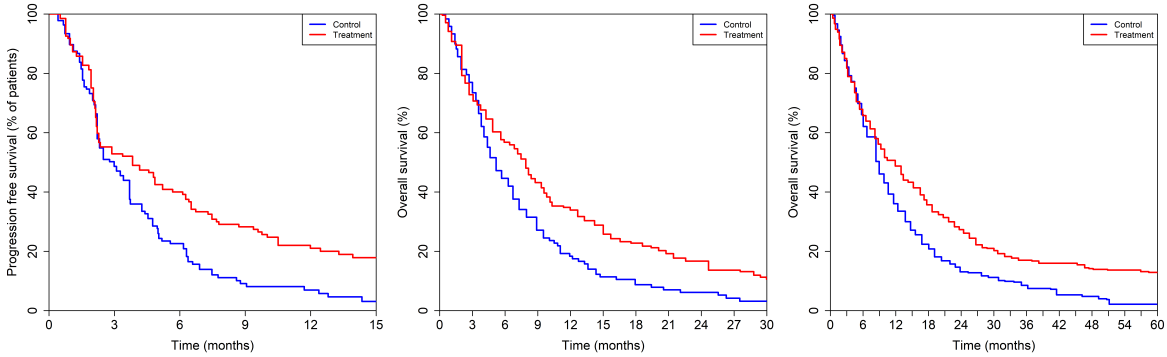


Figure 4.10: *Reconstructed Kaplan–Meier curves for three immuno-oncology trials exhibiting delayed treatment effects: (a) CheckMate 017 (Brahmer, Reckamp, et al., 2015), (b) CheckMate 141 (Yen et al., 2020), and (c) CheckMate 017 + 057 combined (Borghaei, Gettinger, et al., 2021).*

Parameter estimation for each method is summarised in Table 4.3. For both methods, control-arm parameters (λ_c, γ_c) were obtained using `survreg(dist="weibull")`, and

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

Table 4.3: *Parameter estimation under Methods A and B. Method A fixes $\gamma_e = \gamma_c$; Method B estimates γ_e freely. MLE = Maximum Likelihood Estimation.*

τ was estimated visually from the Kaplan–Meier plots. Remaining treatment-arm parameters were then fitted by maximum likelihood estimation (MLE).

Table 4.4 shows the fitted parameters for the CheckMate 017 trial. The corresponding parametric survival curves for both methods are plotted in Figure 4.11.

Method	λ_c	γ_c	τ	γ_e	λ_e
A	0.22	1.29	3	1.29	0.097
B				0.678	0.161

Table 4.4: *Estimated parameters for CheckMate 017 (Brahmer, Reckamp, et al., 2015) under both parameterisations.*

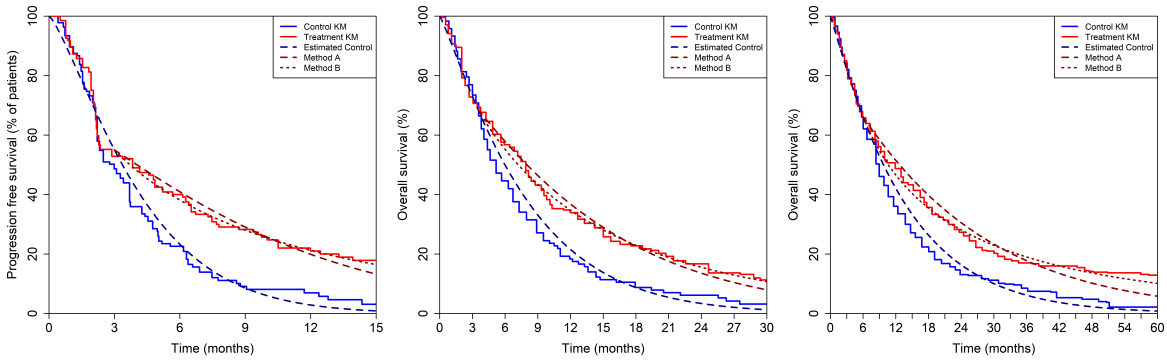


Figure 4.11: *Reconstructed Kaplan–Meier curves with fitted parametric survival curves for both methods. Blue lines show the control fits; red lines show treatment fits from Method A (simplified, solid) and Method B (full, dashed). Despite Method B's extra flexibility, both methods approximate the data closely.*

Although Method B offers a slightly improved fit, unsurprising given its additional free parameter, the difference is minimal. To quantify this, we calculated the power for a range of given sample sizes, assuming the data come from the fitted parameters for each dataset. The resulting power curves are shown in Figure 4.12. We see that

the two curves (one for each method) are nearly indistinguishable, indicating that, for these examples, the simplification $\gamma_e = \gamma_c$ has negligible influence on trial operating characteristics.

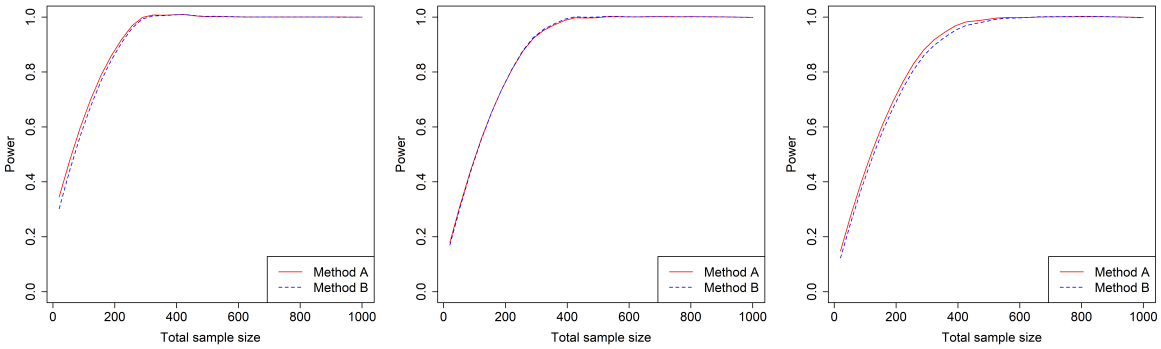


Figure 4.12: *Power curves for both parameterisations (Method A and Method B) across the three datasets. The curves are almost identical, suggesting that the simplification $\gamma_e = \gamma_c$ has no meaningful impact on design conclusions in these examples.*

4.8.2 A more flexible approach to evaluating assurance

While the simplification appears reasonable for the examples above, it constrains the shape of the experimental arm survival curve to be parallel to that of the control arm. Figures 4.13 compares ten individual sampled treatment-arm survival curves and the corresponding pointwise intervals under the simplified (Algorithm 5) and flexible (Algorithm 6) approaches.To address this limitation, we propose an alternative approach, Algorithm 6, which allows independent sampling of the treatment arm shape and scale parameters and thereby enables a wider range of survival curve shapes.

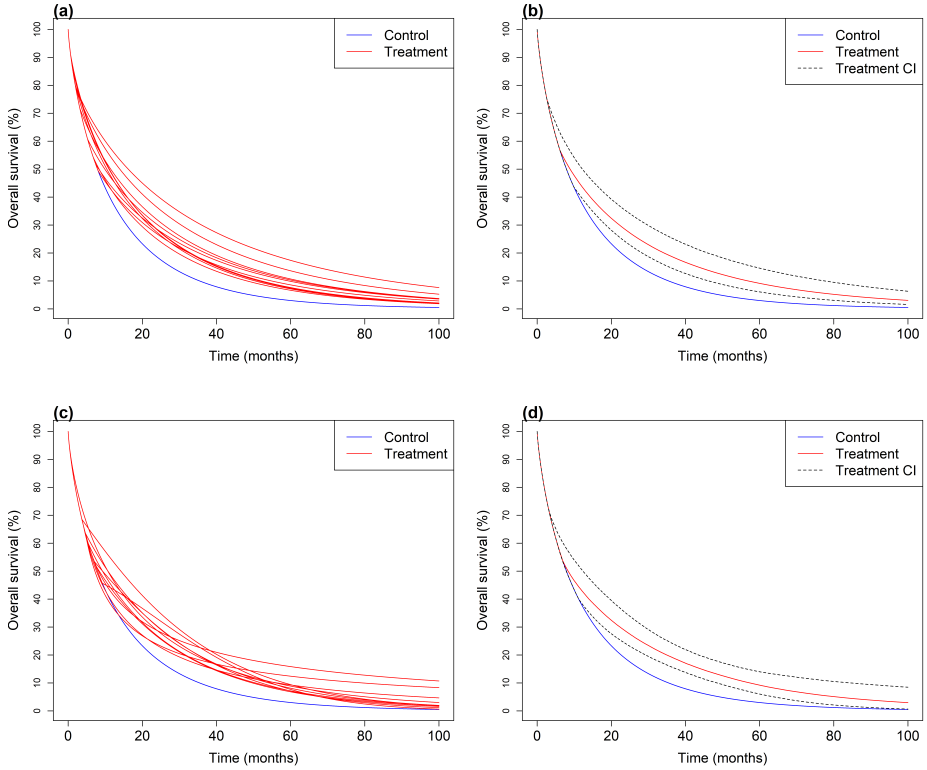


Figure 4.13: *Comparison of sampled treatment-arm survival curves under the simplified (top) and flexible (bottom) approaches. The left panels show ten individual sampled curves, illustrating the range of shapes produced by each method. The right panels show the pointwise 10th and 90th percentiles from 500 sampled curves. While the flexible approach produces more varied individual curve shapes, the resulting pointwise intervals are nearly identical, suggesting that the simplification does not materially affect the overall uncertainty.*

Algorithm 6 Flexible assurance using prior-predictive survival curves

Inputs: Per-arm sample sizes n_c, n_e; priors $\pi(\boldsymbol{\theta_c})$, $\pi(\tau)$, $\pi(\text{HR}^*)$; time grid $\texttt{time} = \{0, 0.01, \ldots, t_{\max}\}$; recruitment model $\mathcal{R}$; censoring model $\mathcal{C}$; primary analysis A; number of prior predictive samples M; number of simulation replicates N

Initialise matrix $\mathbf{P} \in \mathbb{R}^{M \times |\texttt{time}|}$

Step 1: Prior predictive sampling

for $j = 1, \ldots, M$ **do**

  Sample $\boldsymbol{\theta}_{c,j} \sim \pi(\boldsymbol{\theta_c})$, $\tau_j \sim \pi(\tau)$, and $\text{HR}^*_j \sim \pi(\text{HR}^*)$

  Compute treatment survival probabilities on `time` and store in row j of $\mathbf{P}$

end for

Step 2: Assurance simulation

for $i = 1, \ldots, N$ **do**

  Sample $\boldsymbol{\theta}_{c,i} \sim \pi(\boldsymbol{\theta_c})$

  Compute $t_{\max,i}$ as the time at which control survival first falls below 0.01

  Sample $s_{1,i}$ from column of $\mathbf{P}$ corresponding to time $0.25\, t_{\max,i}$

  Sample $s_{2,i}$ from column of $\mathbf{P}$ corresponding to time $0.6\, t_{\max,i}$, ensuring $s_{2,i} < s_{1,i}$

  Sample $\tau_i \sim \pi(\tau)$

  Obtain $(\lambda_{e,i}, \gamma_{e,i})$ via least squares, fitting to $(s_{1,i}, s_{2,i})$

  Generate control and treatment event times using $\boldsymbol{\theta}_{c,i}$, τ_i, $\lambda_{e,i}$, $\gamma_{e,i}$

  Apply recruitment model $\mathcal{R}$ and censoring model $\mathcal{C}$

  Perform primary analysis A and set $U_i = \mathbf{1}(\text{analysis successful})$

end for

Estimate assurance, where R denotes the event of a successful analysis:

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

Since the Weibull survival function is nonlinear in its parameters, the least squares criterion is not guaranteed to be convex, and multiple local minima may exist. Uniqueness of the solution therefore cannot be guaranteed in general; in practice, sensible starting values are used to ensure the solution found is reasonable.

Algorithm 6 proceeds in two steps. In Step 1, a prior predictive sample of M treatment arm survival curves is generated. For each sample j, control parameters $\boldsymbol{\theta}_{c,j}$, a delay time τ_j, and a post-delay hazard ratio HR^*_j are drawn from their respective priors, and the corresponding treatment survival probabilities are computed on a fine time grid and stored in a matrix $\mathbf{P}$. This matrix captures the full range of plausible treatment arm survival trajectories implied by the prior.

In Step 2, the assurance is estimated via simulation. For each replicate i, control

parameters $\boldsymbol{\theta}_{c,i}$ are sampled and used to determine the effective follow-up time $t_{\max,i}$, defined as the time at which the control survival probability first falls below 0.01. Two survival probabilities, $s_{1,i}$ and $s_{2,i}$, are then sampled from the columns of $\mathbf{P}$ corresponding to times $0.25\,t_{\max,i}$ and $0.6\,t_{\max,i}$ respectively, with the constraint that $s_{2,i} < s_{1,i}$. A delay time τ_i is drawn from its prior, and Weibull parameters $(\lambda_{e,i}, \gamma_{e,i})$ are obtained by solving for the values that match the sampled survival probabilities at the two landmark times. The resulting treatment arm survival curve is then used to simulate event times, and the primary analysis is performed as in Algorithm 5.

The procedure for generating a single treatment survival curve under this flexible framework is depicted in Figure 4.14. A delay time τ is drawn from its prior (panel a). Two survival probabilities, s_1 and s_2, are then sampled at fractions $0.25\,t_{\max}$ and $0.6\,t_{\max}$ of the follow-up time (panels b–c). A Weibull distribution is fitted through these points using least squares (panel d). Adjusting these time fractions would allow more or less restrictive flexibility, depending on the desired range of plausible shapes.

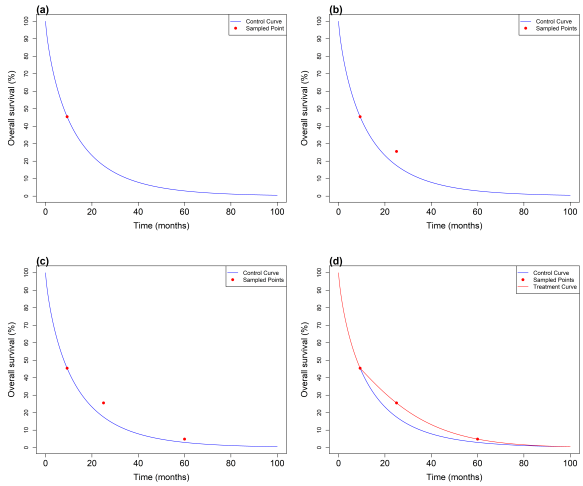


Figure 4.14: *Illustration of the flexible assurance process. (a) A delay time τ is drawn from its prior; (b–c) two survival probabilities s_1 and s_2 are sampled at times $0.25\,t_{\max}$ and $0.6\,t_{\max}$ respectively; (d) a piecewise Weibull model is fitted through these points via least squares* to define the treatment-arm survival curve.

The resulting sampled treatment curves can then be used within the same assurance-calculation framework as before. In practice, the two approaches produce comparable overall uncertainty ranges, but the flexible version relaxes the constraint that treatment curves must mirror the shape of the control curve.

Several alternative approaches to fitting the treatment arm parameters are possible. If survival probabilities were sampled at more than two anchor times, the system would become overdetermined, and one could instead minimise the integrated squared

difference between the fitted and target survival curves over $[0, t_{\max}]$, or maximise a likelihood based on simulated event times. These approaches would offer greater global accuracy at the cost of additional complexity and computational burden. The choice of two anchor points is deliberate: it matches the number of free parameters $(\lambda_{e,i}, \gamma_{e,i})$, is computationally efficient within the simulation loop, and avoids the need for a more complex optimisation procedure.

4.9 Summary

This chapter has presented a methodological framework for calculating assurance in clinical trials where delayed treatment effects (DTEs) are anticipated. Such effects, increasingly common in immuno-oncology and other therapeutic areas involving time-varying mechanisms of action, challenge the traditional assumption of proportional hazards that underpins most survival trial designs. By explicitly modelling the delay between treatment initiation and onset of therapeutic benefit, the framework developed here allows for a more realistic and transparent quantification of trial success probabilities.

We began by describing how survival models can be parameterised to capture DTEs through a piecewise specification of the hazard function. Both exponential and Weibull models were considered, balancing flexibility with interpretability. The exponential model provided a straightforward starting point, while the Weibull model enabled greater adaptability through its shape parameter. Recognising the practical challenges involved in eliciting beliefs about multiple shape and scale parameters, we introduced a simplification that assumed equal shape parameters between arms ($\gamma_e = \gamma_c$). This allowed experts to express beliefs about observable quantities such as the post-delay hazard ratio, rather than abstract distributional parameters, thereby facilitating the elicitation process without introducing undue cognitive burden.

A key contribution of this chapter is demonstrating how expert elicitation can be integrated directly into assurance-based trial design. By eliciting expert beliefs about the likelihood of curve separation, the probability and duration of a delay, and the magnitude of the post-delay hazard ratio, we construct prior distributions that formally encode uncertainty in each component of the delayed treatment effect. This hierarchical elicitation approach ensures that uncertainty is coherently propagated through the assurance calculation within a Bayesian framework. In contrast to conventional power calculations, which typically condition on fixed parameter values, assurance quantifies the probability of trial success under genuine uncertainty, a particularly valuable feature in early or high-risk development programmes.

We further examined the implications of the modelling simplification used in the

Weibull case. Through empirical investigation across several immuno-oncology trials, we found that fixing $\gamma_e = \gamma_c$ led to only negligible differences in both fitted survival curves and resulting power estimates. Nonetheless, we proposed an alternative, more flexible assurance algorithm that removes this constraint, allowing for the generation of treatment survival curves under independent Weibull parameterisation. This provides users with the flexibility to select between computational efficiency and model expressiveness according to the needs of their application.

Two illustrative examples were provided, one using exponential and one using Weibull parameterisations, to demonstrate the practical implementation of the proposed framework. These examples showed how elicited priors and historical data can be combined to calculate assurance under a variety of design assumptions, including sample size, recruitment period, and event-driven stopping rules. The accompanying open-source R package, `DTEAssurance`, offers an accessible and reproducible means of performing these analyses, integrating elicitation, prior construction, and assurance computation into a workflow.

Taken together, the methods developed in this chapter contribute to a broader shift towards more evidence-informed and uncertainty-aware trial design. They provide a practical bridge between statistical methodology and clinical expertise, allowing decision-makers to account for plausible deviations from proportional hazards and to assess design robustness under realistic assumptions.

While the present framework focuses on fixed trial designs, many of the principles extend naturally to adaptive settings, where interim information may be used to update beliefs or modify the course of the trial. The next chapter introduces adaptive designs more formally, with Chapter 6 extending the work in this Chapter, exploring how assurance can be applied to adaptive trial designs that accommodate delayed treatment effects, thereby further enhancing flexibility and decision-making efficiency in complex clinical development programmes.

Chapter 5

Adaptive Clinical Trials

5.1 Introduction

Modern biomedical development increasingly involves complex disease mechanisms, targeted therapies, and heterogeneous patient populations. Within this landscape, traditional fixed clinical trial designs can be challenging to implement efficiently. Sponsors face rising development costs, recruitment difficulties, and extended timelines, all of which place pressure on trial designs to make more effective use of accumulating information.

Randomised controlled trials (RCTs) remain the gold standard for generating reliable evidence, but when implemented with a fixed design structure they provide limited flexibility to modify aspects of the study in response to emerging data. Adaptive RCTs address this limitation by allowing pre-specified changes, such as early stopping, sample size re-estimation, or arm dropping, while maintaining the integrity and validity of the trial (Berry, 2006; Pallmann et al., 2018). In settings where uncertainty is high or development timelines are constrained, such adaptations may offer efficiency gains or more informative decision-making compared with strictly fixed designs. Systematic reviews of published adaptive trials indicate that oncology is the most common therapeutic area in which adaptive designs have been implemented, reflecting both the prevalence of time-to-event endpoints and the substantial uncertainty that characterises late-phase oncology development (Ben-Eltriki et al., 2024).

The development of adaptive trial methodology reflects two complementary historical strands. Early Bayesian work, such as the response-adaptive allocation proposed by Thompson, 1933, introduced the idea that accumulating data could be used to modify trial conduct. In parallel, advances in sequential analysis led to formal procedures for early stopping based on interim data, culminating in the introduction of group

sequential designs by Pocock, 1977 and O'Brien and Fleming, 1979. These methods provided a structured framework for interim analyses while preserving frequentist error control, and were subsequently extended to time-to-event endpoints (Tsiatis, 1981; Tsiatis, 1982). Together, these developments established the statistical foundations for many adaptive designs used in confirmatory clinical trials today.

Adaptive designs (ADs) provide a structured framework for incorporating such pre-planned modifications into an ongoing randomised trial. Adaptations may include early stopping for efficacy or futility, sample size re-assessment, or changes to the randomisation ratio, all implemented on the basis of interim data and under prespecified decision rules. When correctly specified, these designs preserve the integrity of the trial and the validity of its statistical inference (Mehta and Pocock, 2011; U.S. Food and Drug Administration, 2019). From a methodological perspective, adaptive designs for confirmatory trials are supported by a well-developed statistical theory, including group sequential methods, conditional error approaches, and combination testing frameworks, which ensures valid inference following adaptation (Bretz et al., 2009). Complementary regulatory guidance emphasises that adaptive designs, when prospectively planned and rigorously implemented, represent a principled extension of conventional fixed designs rather than an ad hoc departure from them (Bhatt and Mehta, 2016).

Adaptive designs do not, however, eliminate uncertainty. Each adaptation rule defines a set of operating characteristics and can influence power, Type I error, and the interpretation of the final analysis. Regardless of whether a design is adaptive or fixed, it is therefore essential to quantify the probability that the specified trial will achieve its objectives under realistic pre-trial uncertainty about the treatment effect. Assurance, as introduced in Chapter 2, provides a principled framework for this evaluation. It does not alter the design or its adaptation rules; instead, it quantifies the probability of success for the design as specified, averaging over uncertainty in the underlying parameters.

This chapter introduces the core principles underlying adaptive clinical trials. We describe the statistical concepts of conditional power and predictive probability that motivate interim decision-making, and present group sequential designs as one of the earliest and most widely adopted forms of adaptation in confirmatory research. Adaptive design methodology spans a broad spectrum, from dose-finding procedures and response-adaptive randomisation to fully model-based Bayesian approaches. A comprehensive review is beyond the scope of this chapter. Instead, the focus is on conditional power, predictive probability, and group sequential designs, which exemplify the probabilistic reasoning underlying most adaptations used in late-phase trials. These approaches are emphasised for three reasons. First, they provide a transparent basis for interim decisions by linking accumulating data to well-defined operating characteristics. Second, they represent the most established and regulatory-recognised forms of adaptation in confirmatory settings. Third, they form the methodological foun-

dation for the following chapter, where prior uncertainty about treatment effects is incorporated into the evaluation of adaptive designs.

5.2 Aims of the Chapter

The aims of this chapter are to:

- Introduce adaptive designs and explain their role in improving the efficiency and flexibility of modern clinical trials.
- Present key statistical concepts underlying interim decision-making, including conditional power and predictive probability.
- Describe group sequential designs as a foundational framework for planned adaptations.
- Discuss practical and regulatory considerations in the implementation of adaptive methods.

5.3 Conditional Power and Predictive Probability

Interim monitoring is an established component of modern clinical trial design, providing a structured means of evaluating emerging data and determining whether a study should continue, stop early, or be modified. Two quantities commonly used for this purpose are *Conditional Power* (CP) and the *Predictive Probability* (PP) of eventual trial success. Both quantify the likelihood that the final analysis will reject the null hypothesis, but they differ fundamentally in how they treat uncertainty about future outcomes.

Conditional Power is calculated under a fixed assumed value for the future treatment effect. Predictive Probability, in contrast, integrates over uncertainty in that effect using a prior or posterior distribution. Together, CP and PP provide complementary perspectives for interim decision-making, particularly when assessing futility or evaluating whether continued recruitment is justified.

The table highlights the parallel structure between design-stage and interim-stage probability measures.

Stage	Conditioned on Assumed Effect	Integrating Over Uncertainty
Design Stage	Power	Assurance
Interim Stage	Conditional Power (CP)	Predictive Probability (PP)

Table 5.1: *Conceptual relationship between design-stage and interim-stage probability measures. Power and CP are evaluated under a fixed assumed treatment effect, while assurance and PP integrate over uncertainty using a prior or posterior distribution.*

5.3.1 Conditional Power

Conditional Power (CP) is defined as the probability that a study will reject the null hypothesis given the data observed so far and an assumption about the future treatment effect (Jennison and Turnbull, 2000). It is commonly used to assess futility at interim analyses.

Let B_t denote the interim test statistic at information fraction $t \in (0, 1)$, and let θ_f be the assumed true treatment effect for the remainder of the trial. We assume a one-parameter exponential family model under which the test statistic B_t follows a standard normal distribution with mean $\theta\sqrt{t}$ and unit variance, corresponding to the usual large-sample approximation for Wald-type Z-tests. The information fraction t is defined as the ratio of the Fisher information accumulated at the interim analysis to that planned at the final analysis, $t = I_t/I_1$. For time-to-event outcomes, information is typically proportional to the number of observed events, whereas for continuous or binary endpoints it is proportional to the number of accrued observations.

Then the conditional power is given by:

$$\mathrm{CP}_{\theta_f}(t) = \Pr\big(\text{Reject } H_0 \mid B_t,\, \theta_f\big) = 1 - \Phi\left(\frac{Z_\alpha - B_t - \theta_f\sqrt{t}}{\sqrt{1-t}}\right), \tag{5.1}$$

where Φ is the standard normal cumulative distribution function and Z_α is the critical value for a one-sided test at significance level α.

A typical futility rule is to stop the trial if

$$\mathrm{CP}_{\theta_f}(t) \leq 1 - \omega,$$

where ω represents the target power (often 0.8). Thresholds for futility (e.g. CP below 10–30%) are chosen to balance ethical and operational considerations (Walter et al., 2020).

Common choices for θ_f include:

- $\theta_f = 0$: assumes no future treatment effect (the null hypothesis);
- $\theta_f = \hat{\theta}_I$: uses the effect size observed at interim;
- $\theta_f = \theta_\delta$: assumes the effect specified in the original design.

The choice of θ_f is critical: small differences in its specification can lead to large differences in CP and, consequently, in interim decisions. As CP conditions on a fixed future effect, it does not reflect uncertainty in θ. This motivates a Bayesian alternative, Predictive Probability, which integrates over the posterior uncertainty.

5.3.2 Predictive Probability

Predictive Probability (PP) extends the concept of CP by averaging over uncertainty in the true treatment effect (Spiegelhalter, Freedman, and Blackburn, 1986; DeMets, 2006). It represents the probability, given the interim data, that the trial will ultimately demonstrate statistical success.

As with assurance (Section 2.3), the terminology surrounding PP has been inconsistent across the literature. Related terms include: 'predictive power' (Dmitrienko and Wang, 2006; Lan, Hu, et al., 2009), 'Bayesian predictive power (BPP)' (Rufibach et al., 2016), 'predictive probability of statistical significance' (Saville et al., 2014), 'probability of study success' (Wang, Fu, et al., 2013), and 'predictive probability of success (PPoS)' (Kundu et al., 2023). All refer to the same underlying quantity: the posterior predictive probability that the final analysis will reject the null hypothesis.

For clarity and consistency, throughout this thesis we use the term *Predictive Probability (PP)*:

$$\mathrm{PP}(t) = \int \mathrm{CP}_\theta(t)\; p(\theta \mid \mathrm{Data}_t)\, d\theta, \tag{5.2}$$

where $p(\theta \mid \mathrm{Data}_t)$ is the interim posterior distribution of the treatment effect. Conceptually, PP can be viewed as the posterior predictive analogue of assurance—the probability of success conditional on the data observed so far.

A simple futility rule is:

$$\mathrm{PP}(t) < \tau,$$

where τ is typically chosen between 0.05 and 0.20.

	Control ($n = 63$)	Treatment ($n = 78$)
Not raised cTnI (desirable)	40 (63%)	47 (60%)
Raised cTnI (undesirable)	23 (37%)	31 (40%)

Table 5.2: *Interim results from the Moxonidine trial. Raised cTnI indicates higher levels of myocardial ischaemia.*

5.3.3 Illustrative Example: The Moxonidine Trial

To illustrate the use of CP and PP, we revisit the Moxonidine trial introduced in Section 2.5.1. This two-arm study evaluated the effect of Moxonidine on myocardial ischaemia, measured using elevated cTnI levels. An interim analysis was planned after approximately one third of the total information had accrued.

5.3.3.1 Interim Data

At the interim analysis, 141 patients had been randomised: 78 to treatment and 63 to control. The observed results are summarised in Table 5.2.

The interim data suggested a negative treatment effect: 37% of patients in the control arm had raised cTnI levels compared with 40% in the treatment arm, where higher rates indicate greater myocardial ischaemia. On this basis, the study was stopped early for futility.

5.3.3.2 Conditional Power

We retrospectively calculate conditional power at this interim analysis. We do this for three different assumptions for the future treatment effect, θ_f:

- **Null effect:** $\theta_f = 0 \Rightarrow \text{CP} = 0.26\%$;
- **Observed effect:** $\theta_f = \hat{\theta}_I = -0.03 \Rightarrow \text{CP} = 0.03\%$;
- **Planned effect:** $\theta_f = \theta_\delta = 0.15 \Rightarrow \text{CP} = 17.8\%$.

The relationship between CP and θ_f is shown in Figure 5.1, highlighting how CP varies with different effect-size assumptions. The uniformly low conditional power values reflect the unfavourable interim data: the treatment arm performed worse than control, resulting in a low probability of eventual success even under optimistic assumptions about the future treatment effect. This pattern illustrates that conditional power

is primarily driven by the evidence accrued at the interim analysis; when interim results are poor, the likelihood of ultimately rejecting the null hypothesis remains small regardless of the assumed effect size.

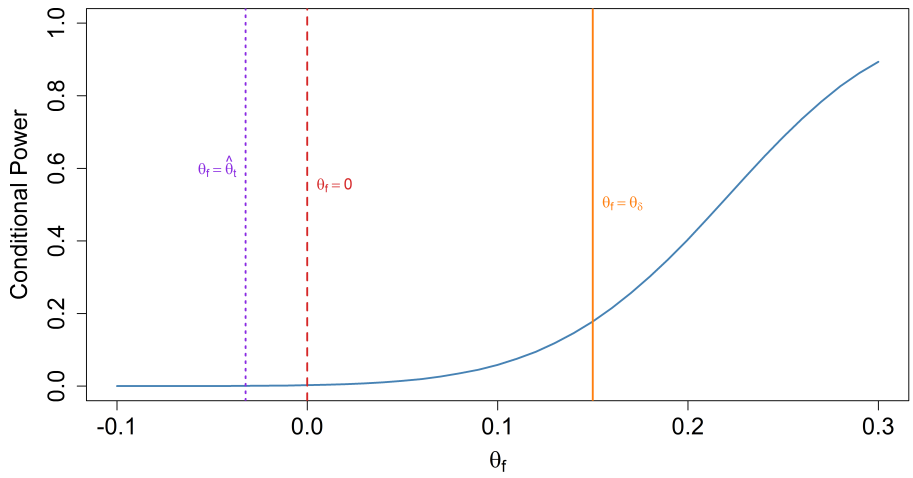


Figure 5.1: *Conditional power as a function of the assumed future treatment effect (θ_f) for the Moxonidine trial. The curve illustrates how the probability of eventually rejecting the null hypothesis depends on the assumed difference in incidence rates between treatment and control. Vertical lines indicate the conditional power under the observed effect ($\theta_f = -0.03$), null effect ($\theta_f = 0$) and planned effect ($\theta_f = 0.15$).*

5.3.3.3 Predictive Probability

Rather than retrospectively evaluating the conditional power at the interim analysis, we can adopt a fully Bayesian perspective and compute the *predictive probability* of eventual trial success. The Predictive Probability (PP) quantifies, given the current data and prior assumptions, the probability that the final analysis will yield a statistically significant result.

Two approaches are available for updating the prior distributions in light of observed data. In the case of conjugate priors, the posterior distributions can be derived analytically. In non-conjugate situations, however, posterior sampling typically requires a Markov Chain Monte Carlo (MCMC) approach. We illustrate both methods below, beginning with the conjugate case.

5.3.3.3.1 Conjugate Case

In the first formulation, we assume independent Beta priors for the treatment and control event rates. This independence implies no prior correlation between arms, allowing conjugate updating and analytic computation of the predictive probability.

We assume the control and treatment group event rates, θ_c and θ_e, each follow independent Beta prior distributions, as detailed in Section 2.5.1:

$$\begin{aligned}\theta_c &\sim \text{Beta}(10.7, 13.1),\\ \theta_e &\sim \text{Beta}(6, 14).\end{aligned}$$

Given binomially distributed data, the Beta–Binomial conjugacy allows us to analytically update these priors. With interim data of $x_c = 23$ events out of $n_c = 63$ for the control arm, and $x_e = 31$ events out of $n_e = 78$ for the treatment arm, the posterior distributions are:

$$\begin{aligned}\theta_c \mid x &\sim \text{Beta}(10.7 + 23,\ 13.1 + 63 - 23) \equiv \text{Beta}(33.7, 53.1),\\ \theta_e \mid x &\sim \text{Beta}(6 + 31,\ 14 + 78 - 31) \equiv \text{Beta}(37, 61).\end{aligned}$$

The prior effective sample sizes (ESS) for the Beta distributions may be used to assess how strongly the interim data influence the posterior. For a Beta(a, b) prior, the ESS is commonly defined as $a + b$ (Morita et al., 2008). This gives an ESS of 23.8 for the control arm and 20 for the treatment arm, compared with interim sample sizes of 63 and 78, respectively. Thus, the observed data contribute substantially more information than the priors, and the posterior distributions are dominated by the interim outcomes. This behaviour is reflected in Figure 5.2: the posterior densities (dashed lines) shift away from the prior means (solid lines) towards the data-driven estimates, resulting in two posterior distributions that are close to each other, consistent with little or no treatment effect at interim.

Using these posteriors, we can compute the Predictive Probability (PP) of ultimately rejecting the null hypothesis in the completed trial. This is achieved by integrating over the joint posterior distributions of θ_c and θ_e (Equation 5.2), and simulating the remaining patients' outcomes to evaluate the probability that the final test statistic will exceed the critical threshold.

For the present interim data, the resulting PP is approximately 2%, indicating a negligible probability of ultimate trial success. This result is intuitive, as the posterior distributions for θ_c and θ_e are nearly overlapping, implying little evidence of a treatment effect.

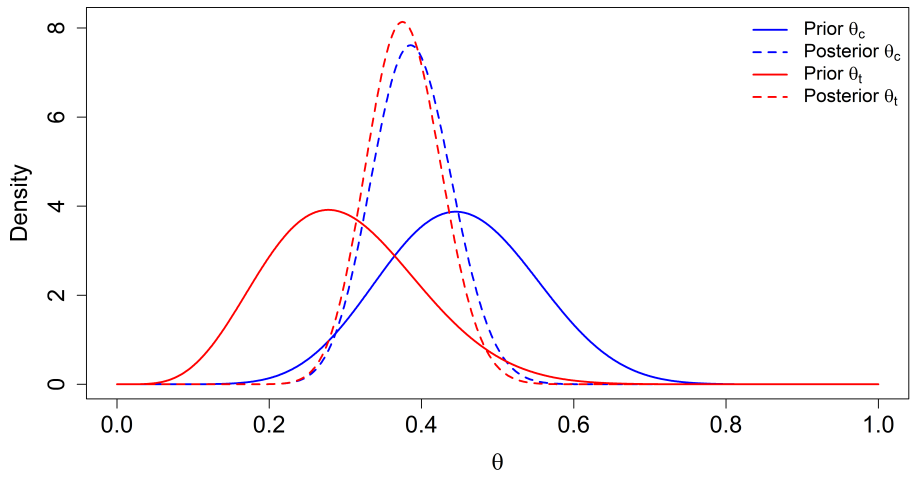


Figure 5.2: *Prior and posterior Beta distributions for θ_c and θ_e under the conjugate Beta–Binomial model. The prior distributions (solid lines) are updated with the interim binomial data to yield the posterior distributions (dashed lines).*

5.3.3.4 Non-Conjugate Case

To relax the independence assumption, we reparameterise the model in terms of the treatment effect $\rho = \theta_c - \theta_e$, which induces dependence between the arm-specific rates. This specification is often more clinically plausible, but it breaks conjugacy: while θ_c may retain a Beta prior, the implied prior for ρ does not yield a conjugate form with the Binomial likelihood. Consequently, closed-form posterior and predictive calculations are no longer available, and numerical methods such as Markov Chain Monte Carlo (MCMC) are required to approximate the joint posterior distribution (Brooks et al., 2011).

We use the same parameterisation as introduced in Section 2.5.1:

$$\theta_c \sim \text{Beta}(10.7, 13.1), \qquad \rho \sim \mathcal{N}(\mu_\rho, \nu),$$

with three prior scenarios reflecting differing levels of informativeness:

- **Scenario 1:** Highly informative prior centred on a 15% treatment effect, $\rho \sim \mathcal{N}(0.15, 0.0001)$.
- **Scenario 2:** Moderately informative prior with greater uncertainty, $\rho \sim \mathcal{N}(0.15, 0.01)$.
- **Scenario 3:** Moderately informative prior centred on a smaller effect, $\rho \sim \mathcal{N}(0.10, 0.01)$.

Figure 5.3 displays the three prior distributions for ρ.

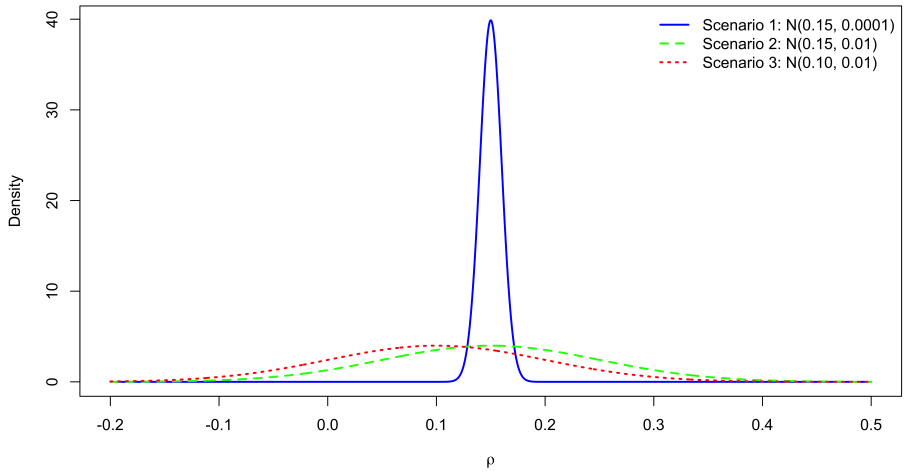


Figure 5.3: *Prior distributions for ρ.*

We update these prior distributions with the interim data to generate posterior distributions, which are shown in Figure 5.4. In the left panel, the first scenario is shown. The prior is largely unaffected by the interim data, and the posterior remains very similar to the prior, reflecting the strength of the prior specification. In the right panel, the second and third scenarios show the prior distributions being pulled towards zero, consistent with the interim data suggesting little evidence of a treatment benefit over control.

In the same manner as in the previous section, we are able to use these distributions to calculate PP. When we do so, we find that in:

- Scenario 1, PP is: 18.3%
- Scenario 2, PP is: 3.7%
- Scenario 3, PP is: 2.7%

These results illustrate how strongly the predictive probability depends on both the strength and the centring of the prior on ρ. In Scenario 1, the highly informative prior centred on a 15% benefit overwhelms the unfavourable interim data, producing an inflated PP ($\approx$ 18%) that is driven largely by prior belief rather than observed evidence.

By contrast, in Scenarios 2 and 3, the weaker priors permit the posterior for ρ to move toward zero, yielding PP values (3.7% and 2.7%) that align closely with both the

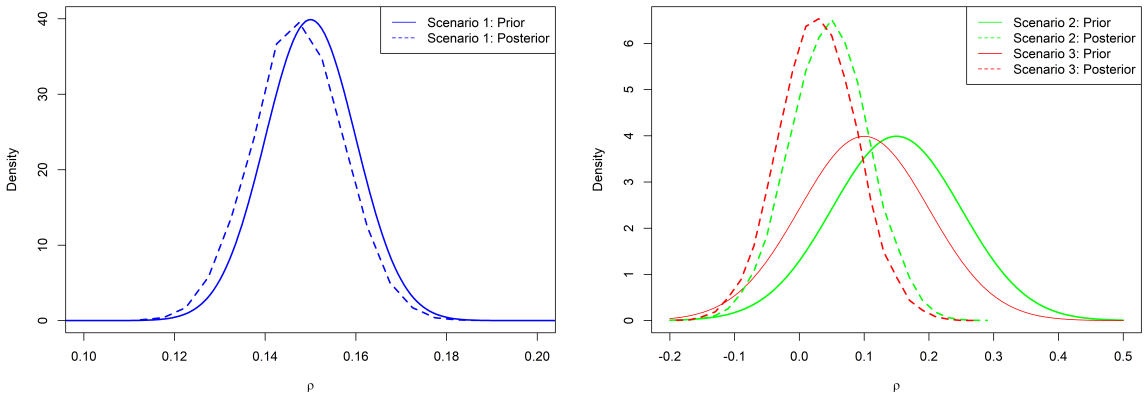


Figure 5.4: *Prior and posterior distributions for ρ for the three scenarios. The first scenario is shown in the left panel and the second and third scenarios in the right panel.*

conjugate-case PP ($\approx 2\%$) and the conditional power results presented earlier. These scenarios demonstrate that when interim data provide little or no evidence of benefit, only very strong or overly optimistic priors can materially raise the perceived chance of eventual success. When priors are calibrated to reflect genuine uncertainty, both CP and PP lead to the same conclusion: the probability that Moxonidine would ultimately demonstrate benefit is extremely low.

It is also noteworthy that the PP under Scenario 1 (18.3%) is almost identical to the conditional power calculated under the planned treatment effect (17.8%). This reflects the fact that a highly informative prior centred on the design assumption effectively forces the predictive probability to behave as if the planned effect were true, despite the contrary interim data.

5.3.4 Summary and Implications

Conditional Power (CP) provides a simple and transparent measure for interim decision-making and remains widely accepted by regulators (Jennison and Turnbull, 2000; Lachin, 2009). Its principal limitation is that it conditions on a fixed assumed treatment effect and therefore does not reflect uncertainty about the true effect size. Predictive Probability (PP), by contrast, integrates over this uncertainty through a posterior distribution, yielding a probability of eventual trial success given both the interim data and prior information (Spiegelhalter, Freedman, and Blackburn, 1986; DeMets, 2006).

PP can be viewed as an "interim assurance": a predictive assessment of whether contin-

uation of the trial is justified. However, its dependence on prior specification introduces sensitivity, especially in confirmatory settings where strong priors may be difficult to defend. Careful elicitation and transparent justification therefore remain essential.

In practice, CP and PP serve complementary roles. CP offers familiarity, ease of communication, and direct alignment with frequentist operating characteristics. PP provides a coherent probabilistic framework that unifies interim monitoring with design-stage reasoning based on assurance. Together, they motivate the need for structured stopping rules that balance statistical validity, efficiency, and ethical considerations.

This motivation leads naturally to Group Sequential Designs (GSDs), which formalise interim monitoring through pre-specified boundaries for stopping for efficacy or futility. The next section introduces GSDs and shows how CP and PP relate to classical sequential testing frameworks.

5.4 Group Sequential Designs

Section 5.3 introduced Conditional Power and Predictive Probability as tools for evaluating emerging evidence at interim. Although these quantities are useful for guiding decisions, they do not, in general, guarantee control of the Type I error rate when multiple interim analyses are considered. In contrast, *Group Sequential Designs* (GSDs) provide a formal framework for interim analyses that preserves the pre-specified frequentist operating characteristics of a trial.

GSDs allow for early stopping for efficacy or futility at a set of planned interim analyses. To control the overall Type I error rate across repeated looks at the data, more stringent boundaries are used at earlier analyses. This error control is commonly implemented through spending functions, which allocate portions of the total α level across the interim and final analyses (Lan and DeMets, 1983; Jennison and Turnbull, 2000). Under this framework, the trial may conclude early only when the accumulated evidence crosses these boundaries, ensuring that repeated testing does not inflate the false-positive rate.

The motivation for adopting GSDs is both ethical and practical. Early stopping for efficacy accelerates the availability of effective treatments, while early stopping for futility prevents unnecessary exposure to ineffective interventions and reduces trial costs. At the same time, GSDs maintain the rigorous error control required in confirmatory research (Jennison and Turnbull, 2000; Proschan et al., 2006).

In practice, GSDs are primarily used to control the Type I error rate when stopping early for efficacy, reflecting the regulatory importance of avoiding inflated false-positive

findings. By contrast, predictive approaches such as CP and PP are more commonly used to guide futility decisions. These decisions are typically viewed as affecting sponsor risk rather than regulatory risk, and therefore tolerate greater flexibility in the probability thresholds employed. As a result, many contemporary adaptive trials combine formal group sequential boundaries for efficacy with predictive or conditional power–based rules for futility.

5.4.1 Statistical Rationale and Error Control

In a group sequential trial, the interim test statistics are evaluated at a sequence of increasing information levels. Under standard regularity conditions, these statistics jointly follow a multivariate normal distribution with correlations determined by the amount of information accrued at each look (Jennison and Turnbull, 2000). Because repeated testing increases the chance of crossing a significance threshold purely by chance, unadjusted interim analyses would inflate the overall Type I error rate (Armitage et al., 1969; O'Brien and Fleming, 1979; Pocock, 1983).

Group sequential methodology addresses this problem by applying more stringent boundaries at earlier analyses, ensuring that the cumulative probability of a false-positive conclusion remains at the pre-specified α level. This framework has been extended to a wide range of endpoints, including survival outcomes (Slud and Wei, 1982).

5.4.2 Construction and Interpretation of Boundaries

At the j-th interim analysis, a pair of stopping boundaries $\{l_j, u_j\}$ is specified. The trial continues if the interim test statistic $Z(t_j)$ lies within these bounds and stops early for futility or efficacy otherwise. Boundaries can be defined on several scales, including standardized Z-statistics, treatment effect estimates, p-values, or Fisher information, depending on what is most interpretable for the data monitoring committee (Gallo, Mao, et al., 2014).

Expressing boundaries in terms of information has been advocated for its monotonicity and ease of interpretation (Whitehead, 1997), whereas effect-size scales may fluctuate more at early looks (Todd et al., 2001). Analytical relationships exist for translating between scales when required (Emerson et al., 2007).

In practice, stopping boundaries are typically constructed using an α-spending function, which allocates the overall Type I error rate across the planned interim analyses. The boundary constants are obtained by numerically evaluating multivariate normal

integrals, ensuring that the joint probability of crossing any stopping boundary under the null hypothesis is controlled at the desired significance level α; standard software such as the `gsDesign` package in `R` (Anderson, 2025) performs these computations automatically. The next sections introduce three widely used spending approaches: Pocock, O'Brien–Fleming, and Wang–Tsiatis.

5.4.2.1 Pocock Boundaries

Pocock (Pocock, 1977) proposed a group sequential procedure in which the stopping boundary for efficacy remains constant across all interim analyses:

$$u_j = C_{\text{Pocock}}(k, \alpha), \qquad j = 1, \ldots, k,$$

where $C_{\text{Pocock}}(k, \alpha)$ is a constant determined by the total number of planned interim analyses k and the overall one-sided significance level α, chosen to control the family-wise Type I error rate at α across all k looks.

This design allows early stopping when the evidence is moderately strong at any look and is easy to communicate and implement. Pocock boundaries tend to spend Type I error more evenly across analyses and therefore allow relatively liberal early stopping compared with more conservative approaches such as O'Brien–Fleming.

A limitation is that Pocock's original formulation is most natural when the number and timing of interim analyses are fixed in advance. Because the same critical value applies at every look, the boundary is less stringent early and more stringent late, which can lead to reduced efficiency if most information accrues near the end of the trial.

5.4.2.2 O'Brien–Fleming Boundaries

O'Brien and Fleming, 1979 proposed boundaries that are highly conservative at early analyses and become progressively less stringent as more information accrues. A common formulation expresses the efficacy boundary at look j as

$$c_j = \frac{C_{\text{OBF}}(k, \alpha)}{\sqrt{t_j}},$$

where t_j is the information fraction at the j-th interim analysis and $C_{\text{OBF}}(k, \alpha)$ is chosen to ensure overall Type I error control.

This structure strongly discourages early stopping when the treatment effect estimate is based on limited data and subject to high variability, reflecting the view that convincing evidence is required before terminating a trial prematurely. As with Pocock's

procedure, the classical O'Brien–Fleming design assumes a prespecified number and timing of interim looks.

The contrasting behaviour of Pocock and O'Brien–Fleming boundaries is illustrated in Figure 5.5. Pocock's boundary remains constant across analyses, whereas the O'Brien–Fleming boundary starts very high and relaxes as information accumulates, providing minimal early error-spending and near-nominal testing at the final look.

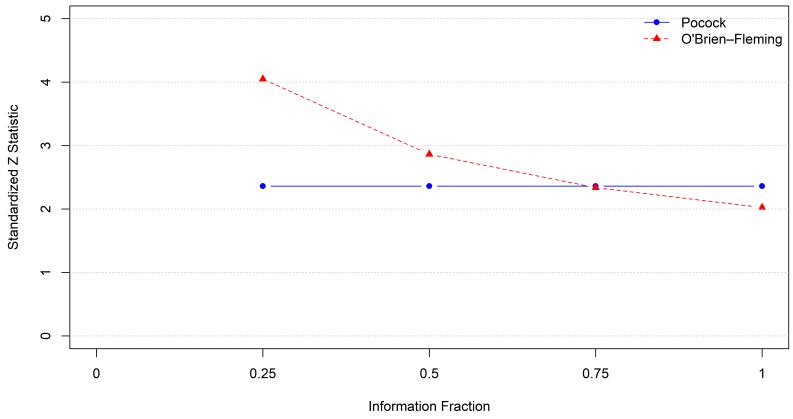


Figure 5.5: *Comparison of Pocock and O'Brien–Fleming stopping boundaries for a one-sided group sequential design with $k = 4$ analyses and $\alpha = 0.025$. Pocock boundaries (horizontal line) require similar evidence at each look, while O'Brien–Fleming boundaries (decreasing curve) demand very strong evidence early and gradually relax over time.*

5.4.2.3 Wang–Tsiatis Boundaries

Wang and Tsiatis, 1987 introduced a flexible parametric family of group sequential designs defined by

$$c_j = C_\Delta(k, \alpha)\, t_j^{\Delta - \frac{1}{2}},$$

where Δ controls the shape of the stopping boundary and $C_\Delta(k, \alpha)$ is chosen to maintain the overall Type I error rate.

This family unifies several classical designs. Setting $\Delta = 0$ yields the O'Brien–Fleming boundary, with strong early conservatism, while $\Delta = 0.5$ approximates Pocock's constant boundary. Intermediate values (e.g., $\Delta = 0.25$) create designs that balance the competing priorities of early stopping flexibility and later-stage power retention.

The Wang–Tsiatis class therefore provides a continuum of options between aggressive

early stopping and highly conservative early boundaries, allowing investigators to tune the operating characteristics to the scientific and ethical aims of the trial.

Figure 5.6 illustrates how varying Δ modulates the trajectory of the stopping boundaries across interim looks.

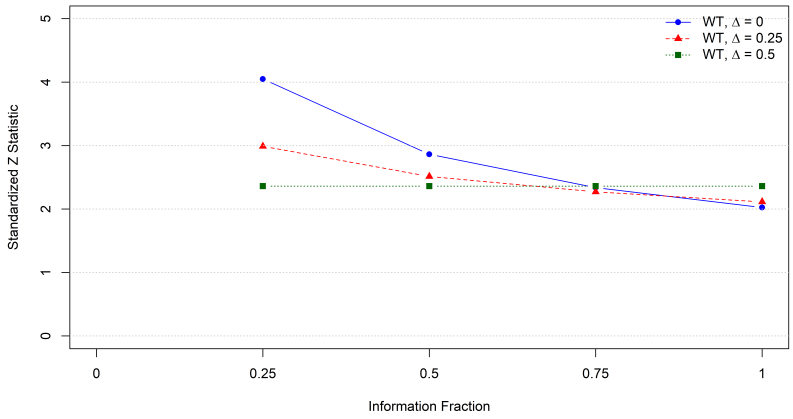


Figure 5.6: *Examples of Wang–Tsiatis stopping boundaries for different values of the shape parameter Δ. Smaller values (e.g., $\Delta = 0$) produce conservative early thresholds akin to O'Brien–Fleming, while larger values (e.g., $\Delta = 0.5$) yield nearly constant Pocock-like boundaries.*

5.4.3 Beta-Spending Functions

The spending functions introduced so far control the Type I error rate (α) across repeated interim analyses and therefore govern early stopping for efficacy. In contrast, early stopping for futility concerns the Type II error rate (β), and classical α-spending approaches do not provide a formal mechanism for regulating this component of the design. To address this limitation, Kim and Demets, 1987 proposed the concept of *dual spending functions*, in which both α and β are allocated over the sequence of interim looks.

In this framework, cumulative error spending up to information fraction t_j is defined through two non-decreasing functions, $\alpha(t)$ and $\beta(t)$, specifying the probabilities of crossing efficacy or futility boundaries by each analysis. The functions are constructed to satisfy the global constraints $\alpha(t_k) = \alpha$ and $\beta(t_k) = \beta$ at the final planned look, ensuring that the design maintains its nominal Type I error rate and power even when early stopping is permitted.

By allocating β across interim analyses, dual spending functions generate futility

boundaries that preserve the planned power under H_1, providing a principled frequentist alternative to predictive or conditional-power methods. However, this flexibility comes with practical challenges. As emphasised by Jennison and Turnbull, 2000, naïve choices of analysis timing or spending schedules can lead to designs that deviate substantially from the intended operating characteristics, requiring careful calibration of information levels and spending functions to avoid unintended inflation or deflation of power.

In practice, formal β-spending is used far less frequently than α-spending. Regulatory guidance places primary emphasis on controlling false-positive risk, while futility stopping is typically viewed as affecting sponsor risk rather than regulatory risk. Consequently, many trials adopt predictive or conditional-power rules for futility, which provide a more direct quantification of the likelihood of eventual success given the interim data. These predictive approaches form the basis for the methodology developed in the next chapter, where we compare their behaviour with β-spending futility in the presence of delayed treatment effects.

5.4.4 The Choice of Stopping Boundaries

Selecting stopping boundaries is a central design decision in a group sequential trial. Different boundary families allocate error differently, influence the probability of early stopping, and affect both the maximum and expected sample size (Pocock, 1977; O'Brien and Fleming, 1979; Jennison and Turnbull, 2000). The choice therefore has direct implications for statistical efficiency and for how interim results are interpreted by monitoring committees.

Key considerations when comparing candidate boundary structures include:

a) flexibility to accommodate changes in the number or timing of interim analyses;

b) the degree of early conservatism;

c) expected and maximum sample size under H_0, H_1, and intermediate effects;

d) probability of early stopping for efficacy or futility; and

e) alignment with the scientific objectives of the trial (e.g., superiority vs. non-inferiority).

Because treatment effect estimates can be unstable early in a trial, boundaries that require strong evidence for early efficacy are generally preferred. Overly permissive rules risk premature conclusions, whereas excessively stringent rules delay termination

and expose participants to ineffective treatments. The challenge is to balance these ethical and statistical considerations.

In practice, α-spending approaches are widely used because they maintain Type I error control while allowing flexibility in the timing of interim looks (Lan and DeMets, 1983; Jennison and Turnbull, 2000). Many spending functions allocate small amounts of α early and increasing amounts later, producing boundaries similar to the O'Brien–Fleming design. Such boundaries are attractive in confirmatory settings where strong evidence is desired before recommending early efficacy. Pocock-type boundaries, which remain constant across looks, simplify interpretation but are less adaptable when the number or timing of analyses may vary.

Boundary symmetry should also reflect the trial objective. Superiority trials typically use symmetric rules for efficacy and futility, whereas non-inferiority or equivalence studies may require stronger evidence to declare non-inferiority than to stop for futility (Proschan et al., 2006; Jennison and Turnbull, 2000). Asymmetric spending rules allow this tailoring.

Although no universally optimal boundary exists, consensus principles include maintaining the planned error rates, achieving adequate power, avoiding excessive early stopping, and keeping expected sample size low under the null. Designs that optimise one criterion, such as minimising expected sample size, may perform poorly on others, particularly in regulatory settings where early stopping must be supported by robust evidence (Fleming, Sharples, et al., 2008).

Ultimately, the selected boundary structure reflects a trade-off between efficiency, ethical considerations, and operational pragmatism. Once chosen, the stopping rule must remain fixed throughout the trial to preserve Type I error control and ensure valid final inference.

5.4.5 Example

Consider a confirmatory two-arm randomised controlled trial with a continuous primary outcome, where larger values indicate better responses. Let $Y_c \sim N(\mu_c, \sigma^2)$ denote outcomes in the control arm and $Y_e \sim N(\mu_e, \sigma^2)$ denote outcomes under the new treatment, with $\sigma = 10$. We assume a clinically relevant effect size $\delta = \mu_e - \mu_c = 5$, corresponding to $\mu_c = 120$ and $\mu_e = 125$.

Under a fixed design, a standard Normal-based power calculation shows that 85 patients per group are required to achieve 90% power to detect $\delta = 5$ at a one-sided Type I error rate of $\alpha = 0.025$.

To explore the implications of implementing a group sequential design, we compare this fixed design with three commonly used boundary systems: Pocock, O'Brien–Fleming (OBF), and Wang–Tsiatis (WT). Each design includes a single interim analysis at 50% information (after approximately 43 patients per arm). The resulting critical Z-values are shown in Table 5.3.

Analysis	Information Fraction	Pocock	OBF	WT ($\Delta = 0.25$)
Interim (Look 1)	0.5	2.178	2.797	2.424
Final (Look 2)	1.0	2.178	1.977	2.038

Table 5.3: *Critical Z-statistic boundaries for group sequential designs with one interim look at 50% information.*

To illustrate the impact on operating characteristics, Table 5.4 reports the expected sample size, power, and probability of early stopping for efficacy under both the null hypothesis ($\delta = 0$) and the alternative ($\delta = 5$). Calculations were obtained using the `gsDesign` package in R.

These results highlight the trade-offs inherent in different sequential monitoring strategies. Pocock's design applies the same boundary at interim and final analyses, placing a relatively large proportion of the Type I error early. This increases the probability of early stopping when the treatment is effective, yielding a substantial reduction in expected sample size (approximately 123 participants under the alternative, compared with 170 in the fixed design). The cost of this efficiency is a modest reduction in power (87% versus the planned 90%) and a higher probability of early rejection under the null, although the overall Type I error remains properly controlled.

The O'Brien–Fleming design is markedly more conservative at the interim look, requiring a large observed effect ($Z \approx 2.80$) for early stopping. As a result, early termination is rare under both hypotheses, and the expected sample size is closer to that of the fixed design. This approach preserves power (approximately 90%) and reduces the chance

True Effect	Design	$P(\text{reject } H_0)$	P(Early Eff.)	$\mathbb{E}[N]$
$\delta = 0$ (null)	Pocock	0.0264	0.0157	168.6
	O'Brien–Fleming	0.0262	0.0033	169.7
	WT ($\Delta = 0.25$)	0.0265	0.0087	169.2
$\delta = 5$ (target)	Pocock	0.8728	0.5470	123.0
	O'Brien–Fleming	0.8990	0.3126	143.1
	WT ($\Delta = 0.25$)	0.8930	0.4525	131.1

Table 5.4: *Operating characteristics of Pocock, O'Brien–Fleming, and Wang–Tsiatis designs under the null and alternative hypotheses.*

of decisions based on unstable early estimates, reflecting its popularity in confirmatory research where robust evidence is required for early success claims.

The Wang–Tsiatis boundary with $\Delta = 0.25$ provides an intermediate option. Its interim boundary is more permissive than OBF but stricter than Pocock, and the final boundary lies between the two. Consequently, its expected sample size and power lie between the corresponding values of the other designs, offering a balanced compromise between efficiency and statistical robustness.

Overall, this example illustrates how the choice of stopping boundary shapes the operating characteristics of a group sequential design. Pocock maximises efficiency but at the cost of reduced power; O'Brien–Fleming preserves power but offers limited opportunity for early stopping; and Wang–Tsiatis provides a flexible middle ground. The appropriate choice depends on the priorities of the trial, such as whether minimising sample size, ensuring high power, or balancing both is the primary aim.

5.4.6 Software

Several commercial and open-source software packages support the design, interim monitoring, and analysis of group sequential trials. Commercially, Cytel's *East* (Cytel, 2025) provides both graphical interfaces and a scripting environment for planning group sequential designs (GSDs), enabling comparisons of stopping probabilities and expected sample sizes under H_0, H_1, and intermediate values.

A number of R packages offer freely accessible functionality:

a) `gsDesign`: Derives group sequential boundaries, computes power and sample size, and produces graphical boundary plots (Anderson, 2025).

b) `rpact`: Implements confirmatory adaptive and group sequential methods, including futility rules, sample-size reassessment, and regulatory-compliant reporting (Wassmer and Pahlke, 2025).

In addition, *SAS* (SAS Institute Inc., 2014) includes two dedicated procedures:

- `PROC SEQDESIGN`: Specifies the group sequential design by defining the number and timing of interim analyses, the boundary family, and the target Type I error and power, and derives the corresponding stopping boundaries and required sample size.

- `PROC SEQTEST`: Implements the specified group sequential design at interim and final analyses, evaluating boundary crossing and providing inference that accounts for sequential monitoring.

Together, these tools allow users to simulate operating characteristics under multiple hypotheses, visualise boundary evolution across information fractions, and generate protocol-ready outputs suitable for regulatory submission.

5.4.7 Summary

Group sequential methods are well established in confirmatory research and are widely accepted by regulators, owing to their rigorous control of Type I error and well-characterised operating characteristics (Jennison and Turnbull, 2000). When compelling interim evidence emerges, they offer ethical and economic advantages by enabling early termination for efficacy or futility. At the same time, when treatment effects are modest, group sequential trials typically continue to their planned maximum sample size and may require a larger maximum N than the corresponding fixed design (Jennison and Turnbull, 2000). Given the generally low success rate of confirmatory trials and the frequent need for extensions, incorporating futility stopping rules at the design stage is usually advisable (Dent and Raftery, 2011).

A range of commercial and open-source software packages, particularly within R, provide support for planning, simulating, and implementing group sequential designs, making their application straightforward in practice.

Successful implementation nonetheless requires careful preparation. The justification for early stopping must be transparent and grounded in clinical, ethical, and operational considerations. Moreover, the feasibility of planned interim analyses depends on predictable information accrual: delays in data collection or processing can complicate the timing of interim looks and limit the practical utility of group sequential monitoring, making these methods more straightforward for short-term endpoints than for time-to-event outcomes (Proschan et al., 2006).

5.5 Adaptive Design Considerations

The use of adaptive features in clinical trials introduces methodological and operational requirements that extend beyond those of fixed designs. Although adaptations can improve efficiency and reduce patient exposure to ineffective treatments, they also increase design complexity and must be justified to regulators, sponsors, and data

monitoring committees (Jennison and Turnbull, 2000; Proschan et al., 2006; U.S. Food and Drug Administration, 2019).

A central requirement is that all planned adaptations, such as sample-size modification, response-adaptive randomisation, or early stopping, are prospectively specified and demonstrated to preserve the trial's statistical validity. This typically requires extensive simulation to show control of the Type I error rate, adequate power, and robustness to plausible deviations from modelling assumptions. Transparent reporting of adaptation rules and their justification is essential to maintain credibility and interpretability (European Medicines Agency, 2007; Bothwell et al., 2018).

Operational feasibility is equally important. Interim analyses depend on timely accrual of high-quality data, reliable estimation of information at each look, and clear separation between trial conduct and interim decision-making (Quinlan and Krams, 2006). Poorly timed or inconsistently executed adaptations can introduce operational bias or compromise the integrity of the study population. Adaptations that alter sample size or recruitment patterns must also be interpreted carefully, as they may influence the precision and generalisability of the final treatment effect estimate (Lehmacher and Wassmer, 1999).

Overall, adaptive designs provide a principled framework for flexible and ethically responsive trials, but their advantages are not automatic. Their successful implementation depends on rigorous statistical planning, operational discipline, and explicit justification of how adaptations support the scientific aims of the study. The following sections examine specific adaptive tools, beginning with group sequential monitoring, and illustrate how these considerations shape their design and use.

5.5.1 Estimands and Interim Analyses

A key principle in adaptive trial design is that the estimand targeted at interim analyses should coincide with the estimand specified for the final analysis. The ICH E9(R1) addendum (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), 2019) emphasises that the estimand defines the clinical question of interest by linking the target population, endpoint, summary measure, and a strategy for handling intercurrent events—that is, post-randomisation events such as treatment discontinuation or use of rescue medication that could affect the interpretation of the endpoint. The estimand should remain stable throughout the study. Recent methodological work, including Collignon et al., 2022, reinforces that inconsistencies between interim and final estimands can undermine both interpretability and the operating characteristics of adaptive designs.

If interim decision rules are constructed around an estimand that differs from the one

used in the final analysis, for example by relying on a short-term surrogate outcome or by defining the treatment effect prior to accounting for intercurrent events, adaptations may unintentionally favour or disadvantage the treatment in ways that are not scientifically coherent. Such misalignment can induce biased stopping behaviour, distort the interpretation of the trial evidence, and conflict with regulatory expectations.

Ensuring a consistent estimand across interim and final analyses therefore helps maintain coherence between the scientific question, the adaptation strategy, and the final inferential framework. When this is not operationally feasible, for instance when a time-to-event estimand depends on follow-up that is not yet available at interim, additional justification is required and the implications for interim decision-making and interpretability must be made explicit.

5.5.2 Timing of Interim Analyses

5.5.2.1 Delay to Primary Outcome

A key practical limitation of group sequential and other adaptive designs arises from the lag between patient enrolment and the availability of primary endpoint data. This delay gives rise to so called *pipeline patients*, who are enrolled but not yet observed at the time of an interim analysis (Sully et al., 2014). When accrual is rapid relative to follow up, as is common in oncology and other survival studies, a large fraction of participants may lack mature data at interim looks. The resulting reduction in effective information can distort stopping decisions and reduce the practical benefit of early analyses (Hampson and Jennison, 2013).

5.5.2.2 Recruitment

The pace of recruitment strongly influences the operational value of interim analyses. If accrual is rapid, most participants may already be enrolled before interim decisions can be implemented, limiting the practical impact of adaptation. Conversely, slower or staggered accrual allows interim results to inform resource allocation, early stopping, or population enrichment, although often at the cost of longer overall study duration. When accrual rates interact with delayed outcome information, as commonly occurs in immuno-oncology, hybrid designs that explicitly account for delayed treatment effects can yield more reliable operating characteristics (Kairalla et al., 2012).

5.5.2.3 Scheduling Interim Analyses

The timing of interim analyses must balance statistical precision with practical utility. Very early looks (below 30% information) are often unstable and increase the risk of premature stopping (Proschan et al., 2006; Edwards et al., 2023). Simulation studies suggest that futility analyses between 50% and 75% of total information often provide the best compromise between efficiency and power (Sully et al., 2014).

Later analyses yield more stable estimates but may delay actionable decisions, particularly when adaptation rules are complex or regulatory consultation is required. Alpha-spending functions allow some flexibility in timing while preserving Type I error control, but data-driven rescheduling must be avoided to prevent inflation of error rates (Whitehead, 1997; Demets and Lan, 1994). Careful pre-specification of timing, decision rules, and communication procedures is therefore essential.

5.5.3 Operational and Logistical Considerations

The successful implementation of adaptive and group sequential designs demands significant operational infrastructure (Gallo, Chuang-Stein, et al., 2006; Quinlan and Krams, 2006; Chow and Corey, 2011). Central to this effort are:

- a detailed interim analysis plan;
- a formal charter for the independent data monitoring committee (IDMC);
- validated standard operating procedures (SOPs) for data collection, cleaning, and locking; and
- clear timelines and communication pathways to ensure independence and confidentiality.

Reliable electronic data capture systems and reproducible interim-reporting workflows are essential (Vandemeulebroecke, 2008; Gaydos et al., 2009). The challenge of pipeline patients again features prominently: when endpoints are delayed relative to recruitment, many participants contribute partial information at interim analyses (Hampson and Jennison, 2013; Whitehead, 1997). Statistical approaches exist to incorporate these contributions formally, but they require careful interpretation to avoid bias.

Equally important is the training of investigators and statisticians. Teams must be capable of running large-scale simulations to evaluate design operating characteristics, implement interim analyses correctly, and ensure that adaptations comply with prespecified rules and regulatory guidance (Chow and Corey, 2011).

5.5.4 Advantages of Adaptive Designs

Adaptive designs offer numerous advantages in clinical research. They improve efficiency by allowing early stopping for efficacy or futility, reducing expected sample size and trial duration without compromising error control (Jennison and Turnbull, 2000; Proschan et al., 2006). From an ethical standpoint, they limit patient exposure to ineffective therapies and speed access to beneficial treatments (Bauer and Köhne, 1994).

Finally, adaptive approaches are increasingly accepted by regulators, provided that all adaptations and decision criteria are pre-specified and transparently reported (U.S. Food and Drug Administration, 2019; European Medicines Agency, 2007). This alignment with regulatory guidance makes adaptive designs a practical, forward-looking strategy for complex modern trials.

5.5.5 Limitations of Adaptive Designs

Despite these advantages, adaptive designs bring challenges in design, analysis, and implementation. They require extensive simulation studies, sophisticated statistical expertise, and precise documentation, increasing planning time and cost (Gallo, Chuang-Stein, et al., 2006; Bauer and Köhne, 1994). Operationally, interim adaptations based on unblinded data pose risks of bias if confidentiality is not strictly maintained (Proschan et al., 2006).

Regulatory scrutiny is stringent, and post hoc modification of adaptation rules can undermine Type I error control and credibility (U.S. Food and Drug Administration, 2019; European Medicines Agency, 2007; Demets and Lan, 1994). Logistically, rapid data turnaround is essential to support timely interim analyses, necessitating dedicated infrastructure and personnel (Vandemeulebroecke, 2008). Inference following adaptation may also be biased, requiring conditional or bias-adjusted estimators (Jennison and Turnbull, 2000).

Overall, adaptive designs demand careful integration of statistical rigour, operational capability, and regulatory compliance. When these components align, they offer a uniquely flexible and efficient framework for evidence generation in modern clinical development.

5.6 Role of Prior Distributions in Adaptive Trial Design

Adaptive designs increasingly rely on quantitative representations of prior knowledge, particularly when uncertainty about control rates, treatment effects, or other design-critical quantities is substantial. In explicitly Bayesian adaptive trials, such information is encoded through prior distributions. However, even in frequentist settings, elicited priors can inform design assumptions, stress-test operating characteristics, and support interim decision-making. This section outlines how prior distributions, whether derived from expert elicitation, historical data, or meta-analysis, contribute to the design, evaluation, and conduct of adaptive clinical trials.

We distinguish three complementary roles for prior distributions in adaptive design:

1. **Evaluating a proposed design:** Priors enable simulation under plausible truths, yielding operating characteristics such as assurance, expected sample size, and probabilities of early stopping.
2. **Optimising design choices:** Priors guide simulation-based comparison across alternative adaptive configurations (e.g., interim timing, futility thresholds, sample-size limits).
3. **Informing interim analyses:** Priors are updated with interim data to form posterior and predictive quantities that support real-time decision-making.

Simulation is central to each of these applications, as analytic results rarely exist for adaptive trials, especially where information accrual is irregular or responses are time-to-event. Later chapters revisit these ideas in the context of delayed treatment effects (DTEs).

5.6.1 Evaluation of a Design

Given a fully specified adaptive design, including sample size, interim timing, allocation ratios, and decision rules, elicited priors allow systematic assessment of performance under uncertainty. This evaluation is conceptually distinct from conventional operating characteristic calculations at fixed parameter values. Rather than assuming a single true treatment effect, the design is assessed with respect to a distribution of plausible values. The resulting prior predictive operating characteristics are directly interpretable for stakeholders, as they reflect current knowledge rather than hypothetical fixed scenarios.

A standard approach is to repeatedly simulate trials by (i) drawing parameters from the prior, (ii) generating datasets under the induced data-generating mechanism, (iii) applying the adaptive rules, and (iv) summarising outcomes such as assurance, expected sample size, expected duration, and probabilities of early stopping for futility or efficacy. These metrics capture how the design performs across the entire range of effects considered credible, and therefore provide a more realistic measure of design adequacy than traditional power calculations.

Regulatory guidance emphasises this type of design evaluation. Both the EMA reflection paper on adaptive designs and the FDA draft guidance state that extensive simulation is required to justify that the trial preserves statistical integrity under a range of plausible conditions. In practice, prior-predictive design evaluation can reveal whether a candidate design is overly conservative, overly aggressive, or sensitive to sources of uncertainty such as accrual patterns, event rates, or delayed onset of treatment effect.

5.6.2 Design Optimisation

Whereas evaluation assesses a single candidate design, optimisation searches across the design space to identify configurations that best meet the trial's objectives. The optimisation problem is naturally framed in terms of a utility function or set of constraints that reflect stakeholder priorities—for example, maximising assurance, minimising expected sample size, or balancing the probabilities of early stopping for futility and efficacy.

Several optimisation strategies are available. Grid search is conceptually simple but can be inefficient when the design space is high dimensional. More sophisticated approaches, such as stochastic search, Gaussian process based Bayesian optimisation, or adaptive regression models, allow efficient exploration of complex spaces in which operating characteristics depend on nonlinear interactions between design elements. In this context, elicited priors serve as the generative distribution for simulation, ensuring that optimisation is anchored to plausible clinical scenarios rather than arbitrary fixed effect sizes.

In confirmatory settings, optimisation must respect constraints required for regulatory validity, including control of the Type I error rate, adequate power, and prespecification of the optimisation procedure itself. This is particularly important when treatment effects may evolve over time or when delayed onset is possible. Optimisation helps identify designs that remain robust across such uncertainties, which cannot be achieved by relying solely on point assumptions. By integrating prior knowledge with simulation-based exploration, optimisation provides a transparent and rational foundation for selecting adaptive design elements.

5.6.3 Using PP as a Decision Rule

Predictive probability (PP) is a widely used framework for futility monitoring in adaptive clinical trials, particularly when early evidence may be inconclusive or when treatment effects are expected to emerge gradually. Across a broad methodological literature, ranging from early Bayesian continuous-monitoring designs (Thall and Simon, 1994) to predictive frameworks for time to event endpoints (Lee and Liu, 2008; Saville et al., 2014; Yi et al., 2012; Jiang et al., 2020), a coherent three step workflow has emerged for operationalising PP based decisions:

1. selecting the timing of the PP analysis,
2. calibrating a futility threshold on the predictive scale,
3. validating the resulting rule through full simulation of operating characteristics.

Unlike conditional power or p-value rules, PP explicitly accounts for uncertainty in both the model parameters and the unobserved future data, aligning interim monitoring with the final analysis estimand.

5.6.3.1 Choosing the Timing of the PP Look

The timing of the PP analysis determines the amount of information available to update the posterior predictive distribution. Very early analyses yield unstable PP estimates, while late analyses allow limited opportunity for efficiency gains. Early Bayesian monitoring frameworks (Thall and Simon, 1994) supported frequent looks, whereas later work emphasised aligning interim timing with the expected emergence of treatment effects, especially in settings with potential delays (Lee and Liu, 2008; Yi et al., 2012; Jiang et al., 2020).

In confirmatory trials that combine Bayesian futility with frequentist efficacy monitoring, a single mid-trial futility analysis is commonly preferred. To select an appropriate timing, we can employ a simulation-based calibration strategy. For a grid of candidate information fractions (e.g., 0.20–0.80), we simulate interim datasets from the prior predictive distribution, update the posterior, and compute the PP. An "informativeness" metric,

$$P(\text{PP} < 0.1 \;\text{ or }\; \text{PP} > 0.9),$$

quantifies how often a given timing yields near-decisive evidence. This mirrors the informativeness criterion in Alhussain and Oakley, 2020. In most settings, information fractions around 40–60% strike an effective balance between reliability and early stopping potential.

5.6.3.2 Calibrating a Futility Threshold

Given a chosen interim timing, a futility threshold c is required such that the trial stops if

$$\text{PP} < c.$$

Threshold calibration proceeds by simulating interim datasets under clinically relevant scenarios (null, delayed-effect, immediate-effect). For each candidate c, we record how often futility is declared incorrectly or correctly. Methods in Thall and Simon, 1994, Lee and Liu, 2008, Saville et al., 2014, and Jiang et al., 2020 generally advocate values in the range 5–20%, balancing false continuation with preservation of power.

5.6.3.3 Embedding PP Within a Group Sequential Efficacy Framework

In confirmatory settings, PP-based futility rules are typically combined with group sequential designs (GSDs) for efficacy. The Bayesian component governs early stopping for lack of benefit, while type I error control is maintained through pre-specified spending functions. Because PP does not account for future efficacy looks within its predictive distribution, the joint behaviour of futility and efficacy rules must be evaluated through forward simulation.

This hybrid Bayesian–frequentist structure is common in oncology (Yi et al., 2012; Saville et al., 2014; Jiang et al., 2020). It is especially advantageous when treatment effects may be delayed: the Bayesian component incorporates delayed-effect models into the predictive distribution, while the frequentist GSD maintains regulatory familiarity.

5.6.4 Summary

Prior distributions provide a principled mechanism for embedding existing knowledge and uncertainty into adaptive trial design. Whether used to evaluate candidate designs, optimise adaptive configurations, or support interim decision-making, priors enable transparent and coherent planning. Their value is greatest when paired with simulation-based evaluation, allowing realistic assessment across plausible scenarios.

5.7 Discussion

Adaptive designs provide a well-validated framework for improving the efficiency, ethical conduct, and informativeness of clinical trials. By enabling early stopping for

efficacy or futility, they can reduce expected sample size, limit unnecessary patient exposure, and accelerate decision-making (Jennison and Turnbull, 2000; Proschan et al., 2006). Their acceptance in regulatory guidance from the FDA and EMA underscores their transition from innovative methodology to established practice in confirmatory research (U.S. Food and Drug Administration, 2019; European Medicines Agency, 2007; Bothwell et al., 2018; Dimairo et al., 2015).

These advantages, however, are contingent on careful implementation. Adaptive designs require rigorous pre-specification, extensive simulation studies, and close operational control (Gallo, Chuang-Stein, et al., 2006; Bauer and Köhne, 1994). Reliable real-time data capture, timely data cleaning, and well-trained data monitoring committees are essential to prevent operational bias and to maintain trial integrity (Quinlan and Krams, 2006; Vandemeulebroecke, 2008). Practical challenges, such as delayed outcomes, uneven accrual, and potential biases introduced by adaptation, remain active areas of methodological development (Lehmacher and Wassmer, 1999).

Consequently, adaptive designs should be viewed as context-dependent tools rather than default choices. They tend to deliver the greatest benefit when treatment-effect uncertainty is substantial, when outcome timing permits meaningful interim analyses, and when adequate logistical infrastructure is available (Chow, Chang, et al., 2005). In settings where these conditions do not hold, the additional complexity may outweigh gains in efficiency.

Overall, adaptive designs broaden the methodological options available for modern clinical trials. Their responsible use requires balancing statistical rigour, operational feasibility, and transparent reporting. The next chapter builds on this foundation by examining how adaptive designs can support decision-making in trials where delayed treatment effects may arise.

Chapter 6

Adaptive Clinical Trial Design with Delayed Treatment Effects [1]

6.1 Introduction

In Chapter 4, we examined assurance calculations for fixed clinical trial designs in the presence of a delayed treatment effect (DTE). Chapter 5 introduced adaptive designs as a framework for improving trial efficiency and ethical conduct through data-driven modifications. Building on both chapters, the present chapter considers adaptive clinical trial design when prior distributions have been elicited for key time-to-event parameters, including the control survival distribution, the duration of delay, and the post-delay treatment effect.

As outlined in Chapter 5, adaptive designs offer clear operational and ethical advantages. However, their application becomes substantially more challenging in settings where treatment effects emerge only after an initial delay period. A substantial body of literature shows that interim analyses can be severely biased when conducted too early, either during periods in which the control and treatment groups have similar outcomes or when the treatment effect has only just begun to develop. In such cases, interim test statistics may be attenuated or may even favour the control arm (Li et al., 2021; Ghosh et al., 2021), leading to violations of proportional hazards assumptions and reduced power for conventional log rank based monitoring procedures. Empirical studies in immuno-oncology further demonstrate that standard futility boundaries, particularly those based on interim hazard ratio estimates evaluated at fixed information times, can prematurely terminate trials with genuine but late emerging benefits, especially under

[1] The research presented in this chapter is currently under review at *Pharmaceutical Statistics* and is available on arXiv.

rapid accrual or longer delay periods (Korn and Freidlin, 2018; Wu et al., 2023). Consequently, prevailing methodological guidance cautions against overly aggressive early stopping in settings characterised by non proportional hazards and delayed separation of survival curves.

In response to these concerns, several strategies that take the delay into account, have been developed. A prominent recommendation, due to Korn and Freidlin, 2018, is that futility analyses should only be undertaken once a substantial proportion of accrued events arise after the anticipated delay period, and that stopping should be reserved for clear evidence of underperformance. Their results suggest negligible loss of power under delayed effects when interim looks are aligned with post-delay information, in contrast to more aggressive hazard-ratio–based rules. Related methodological work reinforces this principle: interim information time should reflect the accumulation of informative (post-delay) events, either by explicitly delaying interim analyses (Wu et al., 2023; Li et al., 2021) or by employing weighted test statistics that down-weight early, non-informative failures.

Complementary approaches address non proportional hazards directly through modified interim and final test statistics. Weighted and combination log rank procedures, including Fleming–Harrington weights, max combo tests, and maximin efficiency robust tests, have been shown to preserve power across a broad range of delayed effect scenarios and have been successfully incorporated into group sequential frameworks without compromising Type I error control (Prior, 2020; Ghosh et al., 2021). Event driven monitoring strategies based on post delay events further reduce the risk of false negative interim conclusions (Wu et al., 2023). Collectively, these developments highlight the importance of careful specification of interim monitoring rules in settings with delayed treatment effects, since failure to account explicitly for the delay can lead to systematically misleading interim conclusions and materially altered operating characteristics.

Motivated by these findings, we evaluate the futility rule proposed by Korn and Freidlin, 2018, which recommends performing futility assessments only once approximately two-thirds of observed events have occurred beyond the expected delay time. We investigate the behaviour of this rule under a range of scenarios to assess its robustness and identify conditions under which it provides reliable guidance for decision-making.

The remainder of the chapter extends the adaptive methodologies introduced in Chapter 5, specifically predictive probability (PP) and group sequential design (GSD), to settings involving delayed treatment effects and parameter uncertainty informed by expert elicitation. We conclude with illustrative examples and demonstrate the software tools implemented in the `DTEAssurance` package developed as part of this research. By integrating elicited priors for delay related parameters directly into both PP and GSD frameworks, this chapter provides the first formulation of predictive probability

for adaptive survival designs with delayed treatment effects, enabling interim decision rules that explicitly account for non proportional hazards.

6.2 Aims of the Chapter

The purpose of this chapter is to investigate how elicited prior distributions on key time-to-event parameters can be incorporated into adaptive clinical trial design when delayed treatment effects are anticipated. The specific objectives are:

- To evaluate an existing futility timing rule for interim analyses, examining its operating characteristics under a range of delay scenarios and identifying conditions under which the rule provides reliable guidance.
- To study Predictive Probability (PP) as a decision-theoretic framework for adaptive interim monitoring in the presence of delayed effects, with particular attention to how prior uncertainty about delay duration and post-delay treatment effect influences predictive assessments.
- To assess assurance for group sequential designs (GSDs) under elicited priors, providing a pre-trial evaluation of power, early stopping probabilities, and overall operating characteristics when non-proportional hazards are expected.
- To demonstrate the proposed method using the `DTEAssurance` R package and associated Shiny application, illustrating practical implementation and facilitating reproducibility.

Through these objectives, the chapter aims to provide both methodological insight and practical tools for adaptive design planning in the presence of delayed treatment effects.

6.3 Investigation of the Korn-Freidlin Proposed Rule

Futility monitoring in time-to-event trials is particularly challenging when treatment effects are expected to emerge only after a delay. Standard interim rules, such as the Wieand rule (Wieand et al., 1994), implicitly assume proportional hazards and therefore risk stopping a trial prematurely if early follow-up contains little or no information about the eventual treatment benefit. Motivated by this concern, Korn and Freidlin,

2018 proposed a simple modification to the Wieand approach intended to guard against inappropriate early futility decisions in settings such as immuno-oncology.

Their method introduces an additional information-timing requirement: an interim analysis is conducted only if at least two-thirds of the observed events have occurred more than three months after randomisation. The three-month threshold reflects empirical observations that immunotherapy effects often manifest after several months of treatment, though Korn and Freidlin, 2018 note that this value can be adapted to reflect the expected delay in other disease and treatment contexts. By ensuring that interim monitoring is based predominantly on post-delay events, the rule aims to reduce the risk of erroneously stopping a study during a period in which the treatment effect is not yet expected to be detectable.

Despite its conceptual appeal and increasing practical interest, the operating characteristics of this rule have been evaluated only in a narrow set of scenarios. In particular, its performance under different recruitment patterns, baseline hazard functions, and degrees of delayed treatment effect remains unclear. Moreover, it is not known whether the rule is overly conservative in some settings or insufficiently protective in others.

In this section, we extend the investigation of Korn and Freidlin, 2018 by examining the behaviour of the rule across a broader class of delay structures and design assumptions. Our goal is to identify the conditions under which the rule provides reliable protection against premature futility stopping, and to highlight situations where its use may lead to undesirable operating characteristics or reduced efficiency.

6.3.1 Trial Setup and Monitoring Rules

The example considered by Korn and Freidlin, 2018 is a two-arm randomised superiority trial enrolling 680 patients (340 per arm) uniformly over 34 months. The primary analysis is based on a one-sided log-rank test at the 2.5% significance level, performed once 512 events have accrued. Survival in the control arm follows an exponential distribution with median 12 months, corresponding to a hazard rate $\lambda_c = \log(2)/12$. Under proportional hazards with hazard ratio (HR) = 0.75, this design yields approximately 90% power.

Korn and Freidlin examined how early futility monitoring performs under a range of survival patterns, reflecting proportional hazards, delayed separation, and crossing hazards. Six generative mechanisms were considered, each evaluated under two administrative follow-up durations (12 and 34 months), yielding 12 scenarios in total:

1. Proportional hazards: HR = 0.75 (beneficial)

2. Proportional hazards: HR = 1.00 (null)
3. Proportional hazards: HR = 1.30 (harmful)
4. Delayed benefit: HR = 1.00 for 3 months, then HR = 0.693
5. Delayed benefit: HR = 1.00 for 6 months, then HR = 0.620
6. Crossing hazards: HR = 1.30 for 3 months, then HR = 0.628

Scenarios 4–6 were calibrated so that, despite their non-proportional hazard structures, the trial maintained approximately 90% power under the assumed design parameters. This ensures meaningful comparability with the proportional-hazards benchmark and isolates the impact of delayed or crossing effects on futility monitoring performance.

For each scenario, four monitoring strategies were evaluated:

- **No Interim Analysis**: The trial proceeds directly to the final analysis at 512 events with no opportunity for early stopping.
- **Wieand Rule**: At 50% and 75% information (256 and 384 events), the trial stops for futility if the observed hazard ratio satisfies HR > 1. i.e., if the interim data indicate worse outcomes on treatment than control.
- **O'Brien–Fleming β-spending Rule** (Jennison and Turnbull, 2000): Non-binding futility boundaries are applied at 33% and 67% information, corresponding to Z-thresholds of 0.011 and 0.864 (approximately HR thresholds of 0.998 and 0.913). Crossing these boundaries triggers a futility recommendation.
- **Proposed Rule (Korn–Freidlin)**: Interim analyses are planned at 50% and 75% information, but an analysis is carried out only if at least two-thirds of the accumulated events have occurred more than 3 months after randomisation. Once an interim look is triggered, futility is declared if the observed hazard ratio satisfies HR > 1, mirroring the Wieand criterion but applied only after sufficient post-delay information has accumulated.

Korn and Freidlin, 2018 found that conventional futility monitoring rules can substantially reduce power in the presence of delayed treatment effects, particularly when accrual is rapid. For the non-delay scenario under which the trial was designed, both the Wieand rule and their proposed modification had negligible impact on power, whereas the more aggressive O'Brien–Fleming approach produced a modest but noticeable reduction. In settings where the treatment was ineffective, all futility rules reduced expected sample size and trial duration as intended. However, under delayed-effect

scenarios, including 3-month and 6-month delays as well as crossing hazards, the performance of the standard monitoring rules deteriorated. With 34-month accrual, the Wieand rule produced small but non-trivial losses of power and the O'Brien–Fleming rule performed poorly, whereas the proposed modification preserved power at essentially the same level as a design with no interim monitoring. When accrual was compressed to 12 months, these issues became more pronounced: both Wieand and O'Brien–Fleming led to unacceptable power loss, with O'Brien–Fleming performing especially badly under crossing hazards. Across all delayed-effect scenarios and accrual patterns, the proposed rule consistently maintained acceptable power while still enabling early stopping for clearly ineffective or harmful treatments.

6.3.2 Investigation

We investigated whether the favourable operating characteristics reported by Korn and Freidlin, 2018 persist more generally, and in particular whether there exist settings in which the proposed rule behaves suboptimally. By construction, the triggering condition for an interim analysis depends on the accrual rate, the baseline control hazard, and the form of the treatment effect. Consequently, the proposed rule can exhibit one of three behaviours:

1. **Wieand**: the two-thirds requirement is satisfied before 50% information, so interim analyses occur as originally scheduled and the rule coincides with the standard Wieand approach;
2. **Delayed Wieand**: the requirement is met after 50% but before 100% information, leading to delayed interim analyses;
3. **No Interim Analysis (No IA)**: the requirement is never satisfied before the final analysis, so no interim looks are conducted.

These three regimes imply simple bounds on the operating characteristics of the proposed rule. Because it can never trigger an interim earlier than Wieand, and can never delay an interim beyond a design with no interim monitoring, its performance with respect to power, trial duration, and expected sample size must satisfy

$$\text{No IA} \geq \text{Proposed} \geq \text{Wieand}.$$

To illustrate these behaviours, Figure 6.1 displays the proportion of events occurring more than 3 months after randomisation as a function of the total number of accrued events under Scenario 1.

In Figure 6.1(a) (recruitment is 34 months), the two-thirds criterion is satisfied at approximately 200 events, which occurs prior to the planned 50% information point

(256 events). In this setting, the proposed rule therefore coincides with the Wieand approach, as the first interim analysis is triggered on schedule.

In contrast, in Figure 6.1(b) (recruitment is 12 months), the criterion is not met until roughly 300 events, i.e., after the 50% information time but before the final analysis. This configuration produces the "delayed Wieand" behaviour, with the first interim analysis deferred until the triggering condition is achieved.

The earlier triggering under slower accrual is a direct consequence of follow-up distribution: with extended recruitment, a larger proportion of enrolled participants have exceeded 3 months of follow-up by the time the event count reaches a given threshold, leading to more rapid accumulation of post-delay events. Under faster accrual, follow-up is more compressed, and a smaller proportion of events arise after 3 months, delaying the point at which the two-thirds requirement is satisfied.

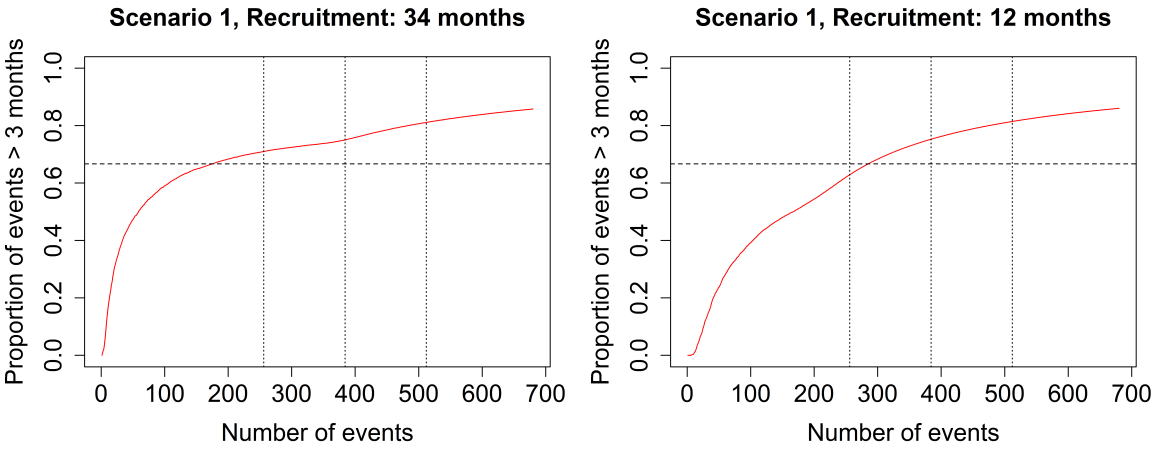


Figure 6.1: *Proportion of events occurring after 3 months versus total events accrued under Scenario 1. (a) Recruitment is 34 months, (b) Recruitment is 12 months. Dashed horizontal line indicates the 2/3 threshold. Vertical lines denote information fractions (50%, 75% and 100%).*

6.3.3 Robustness of the Proposed Rule

Korn and Freidlin, 2018 showed that, across the scenarios studied in their paper, the proposed modification performs well: effective treatments are rarely stopped for futility, and scenarios with delayed benefit trigger appropriately postponed interim analyses that avoid premature negative decisions. Our aim here is not to criticise their rule, which behaves robustly in most practical settings, but rather to identify circumstances in which it may be suboptimal. Understanding these edge cases supports better

decision-making for practitioners who may wish to tailor their monitoring strategy to the expected delay structure and accrual characteristics of their own trial.

Earlier, we observed that the operating characteristics of the proposed rule are always bounded between those of the Wieand rule and the no-interim-analysis (No IA) design. This follows directly from the structure of the triggering mechanism: the proposed rule can never lead to earlier interim analyses than Wieand, nor can it produce analyses later than omitting interim monitoring entirely. Thus, the rule cannot fail catastrophically. However, it can behave suboptimally in specific circumstances. We identify three settings where this occurs:

A: False Stopping: An effective treatment is stopped for futility because early data are misleading.

B: Excessive Delay: The triggering condition is satisfied only after the planned final analysis, so interim futility monitoring never occurs, even when the treatment is clearly ineffective or harmful.

C: False Continuation: Early favourable trends allow an ultimately ineffective treatment to pass interim futility checks, leading to continued enrolment when early stopping would have been desirable.

Although these behaviours arise only in particular regions of the parameter space, there is no single configuration that uniquely induces them. In fact, many combinations of delay lengths, hazard ratios, recruitment patterns, and follow-up durations can produce each of the three behaviours described above. To provide concrete illustrations, Table 6.1 lists representative parameter sets that reliably generate each phenomenon in simulation. These examples are not intended to be exhaustive; rather, they demonstrate the types of conditions under which the proposed rule diverges meaningfully from both the Wieand rule and the No–IA design.

Table 6.1: *Parameters used to generate representative scenarios exhibiting false stopping, excessive delay, and false continuation under the proposed monitoring rule.*

Scenario	Recruitment	λ_c	HR_1	HR_2	Delay
A - Falsely Stopping	0	log(2)/18	1.3	0.65	6
B - Excessive Delay	34	log(2)/6	1.3	1.3	0
C - Falsely Continuing	12	log(2)/15	0.7	1.1	6

6.3.3.1 Scenario A: Falsely Stopping

Scenario A arises under very rapid recruitment and an initially unfavourable treatment effect. As shown in Table 6.1, the control hazard is relatively low (median 18 months), recruitment is extremely fast, and the treatment effect is harmful for the first 6 months before becoming beneficial thereafter.

Figure 6.2 illustrates why this configuration causes the proposed rule to collapse to the Wieand rule. Figure 6.2(a) shows that the two-thirds post–three-months requirement is satisfied at roughly 256 events (the planned 50% information point) so the timing of interims is identical to Wieand. Figure 6.2(b) shows the underlying survival curves. The treatment arm is worse than control during the first 6 months, after which it becomes beneficial; the curves cross at approximately 10 months. The first interim occurs near this crossing point. Because all early data favour the control, the interim hazard ratio is likely to exceed 1, increasing the chance of falsely declaring futility despite the long-term benefit. Thus, even though the proposed rule is intended to protect against premature stopping under delayed benefit, in this configuration it provides no such safeguard.

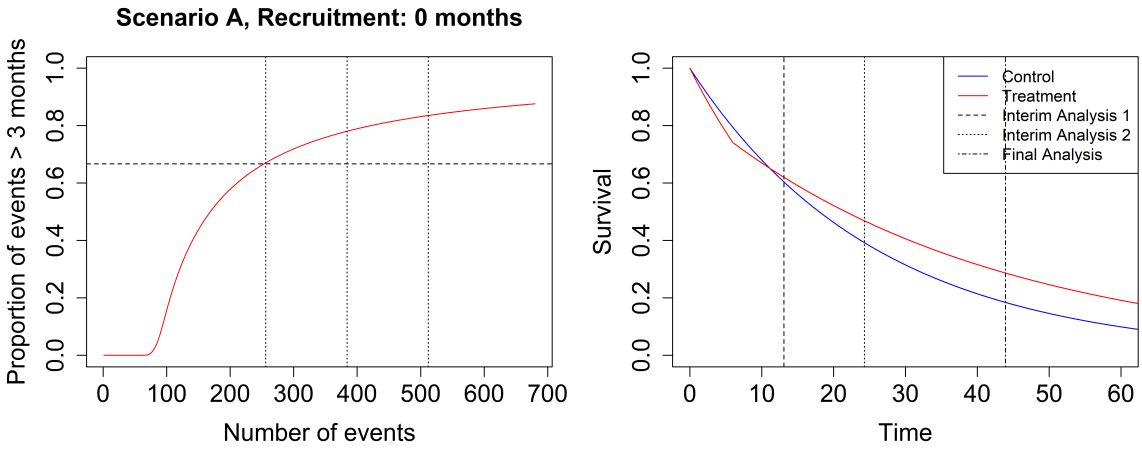


Figure 6.2: ***Scenario A: False Stopping.*** *(a) proportion of events occurring after 3 months versus total accrued events, with the horizontal dashed line marking the 2/3 threshold and vertical lines indicating planned information fractions. The threshold is reached near 256 events, causing the proposed rule to coincide with Wieand. (b) underlying survival curves showing initial harm followed by delayed benefit. The first interim occurs near the point at which the curves cross, making false futility stopping likely.*

6.3.3.2 Scenario B: Excessive Delay

Scenario B arises when the baseline hazard is high (median 6 months), meaning that events accumulate much faster than follow-up time. Because patients experience events early, only a small fraction of observed events exceed 3 months of follow-up, and the two-thirds requirement is never met. The treatment arm is set to be consistently worse than control purely to illustrate the consequences of this delayed triggering.

Figure 6.3 shows that the triggering condition is satisfied only at approximately 650 events, well beyond the 512 events required for the final analysis. Thus no interim analysis is ever conducted under the proposed rule. Although the treatment is harmful, the design behaves identically to the No IA case, missing an opportunity for early stopping that both Wieand and O'Brien–Fleming would have enabled.

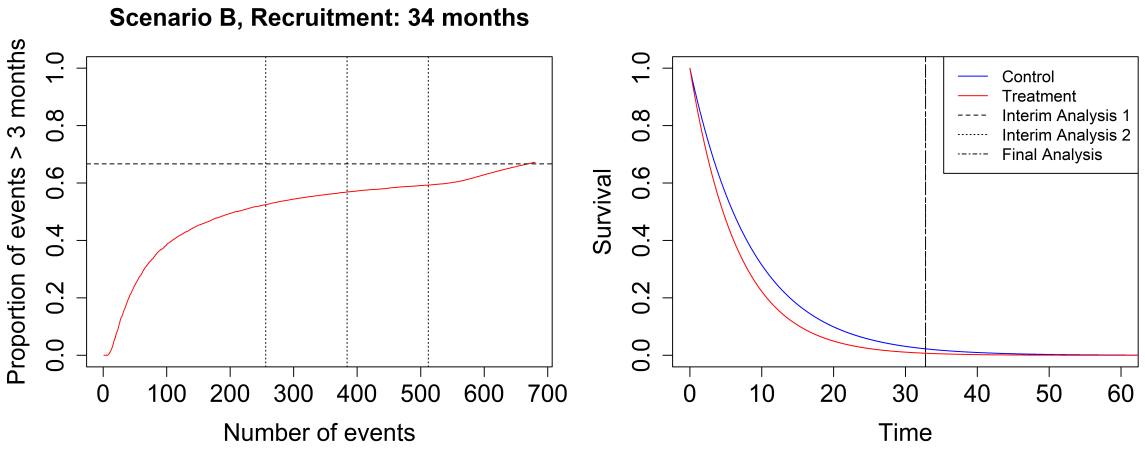


Figure 6.3: ***Scenario B: Excessive Delay.*** *(a) proportion of post–three-month events versus total accrued events. The threshold is met only after approximately 650 events, far beyond the 100% information target of 512 events. (b) survival curves showing consistent harm of the treatment. Vertical lines indicate analysis times; all coincide near 32 months because no interim is ever triggered.*

6.3.3.3 Scenario C: Falsely Continuing

Scenario C arises when the early hazard ratio suggests benefit (HR < 1), but the treatment effect worsens later and ultimately becomes harmful. Because early data appear favourable, the interim hazard ratio does not exceed 1 at the first interim, so the futility boundary is not crossed even though the long-term trend disfavors the treatment.

In this configuration, shown in Figure 6.4, the proposed rule coincides with the Wieand rule; the triggering condition is satisfied before 50% information. The survival curves show sustained early benefit, followed by degradation that makes the treatment arm worse thereafter. The interim analysis occurs while the early benefit still dominates, so the trial continues despite being unlikely to succeed at the final analysis. This scenario illustrates that futility monitoring cannot protect against misleading early over-performance.

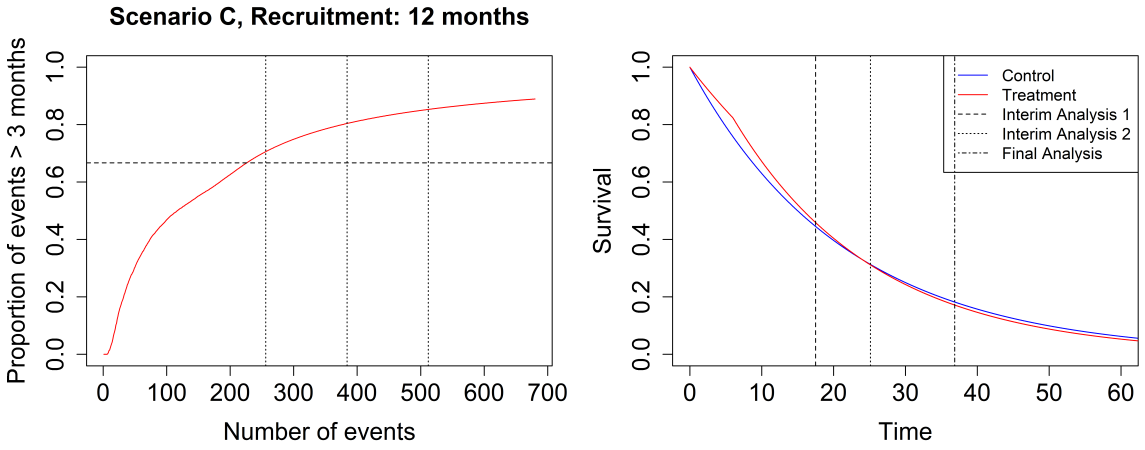


Figure 6.4: ***Scenario C: False Continuation.*** *(a) the triggering condition is satisfied before 256 events, causing the proposed rule to match Wieand. (b) survival curves showing early treatment benefit followed by later harm. The first interim occurs while survival in the treatment arm remains better than control, allowing an ultimately ineffective treatment to pass the futility check.*

6.3.3.4 Results

Table 6.2: *Operating characteristics for failure mode scenarios*

	No Interim Analysis			Wieand Approach			O'Brien-Fleming Approach			Proposed Approach		
Scenario	Power	Duration	SS	Power	Duration	SS	Power	Duration	SS	Power	Duration	SS
A	0.674	42.2	680.0	0.453	27.7	680.0	0.0856	9.9	680.0	0.482	29.2	680.0
B	0	33.0	661.5	0	19.9	398.4	0	15.3	306.5	0	33.0	661.5
C	0.040	36.5	680.0	0.040	32.0	680.0	0.040	28.3	680.0	0.040	32.0	680.0

The numerical results in Table 6.2 highlight how the proposed rule behaves in concrete terms and clarify the practical implications of the three failure scenarios.

6.3.3.4.1 Scenario A: False Stopping

Here the true treatment effect is beneficial, but early events disproportionately favour

the control arm. Under the *No IA* design, power remains reasonably high (0.674), since the trial runs to completion and long-term benefit is eventually observed.

Under the Wieand rule, power drops sharply to 0.453 because the interim hazard ratio frequently exceeds 1, leading to premature futility stopping before the delayed benefit can manifest. The proposed rule also loses power (0.482), but the reduction is less severe: its triggering mechanism delays interim monitoring until more post-delay events are available, partially mitigating the misleading early trend.

By contrast, the O'Brien–Fleming (β-spending) rule performs worst (power ≈ 0.086) because its aggressive early futility boundary makes it highly vulnerable when early data suggest harm.

Interpretation: Scenario A shows that although the proposed rule offers partial protection against premature futility stopping, it cannot fully counteract strongly misleading early data. Severe early imbalance will compromise any futility-monitoring strategy that relies on early event information.

6.3.3.4.2 Scenario B: Excessive Delay

In this scenario, a large proportion of events occur before 3 months, meaning the "two-thirds post–3-month" requirement is not met until after the planned final analysis. Consequently, the proposed rule behaves identically to the *No IA* design, with power $= 0$ and expected sample size ≈ 662.

In contrast, both the Wieand and O'Brien–Fleming rules detect futility early (expected sample sizes ≈ 398 and 306, respectively), correctly terminating the ineffective trial.

Interpretation: Scenario B highlights the clearest vulnerability of the proposed rule: when event times are predominantly early, the triggering condition suppresses futility monitoring entirely, forfeiting the ethical and operational advantages of early stopping.

6.3.3.4.3 Scenario C: False Continuation

Here, early events happen to favour the treatment despite the true effect being null or slightly harmful. All monitoring strategies exhibit similarly low power (≈ 0.04), since the final analysis fails to reject the null. However, trial duration differs:

- O'Brien–Fleming stops earliest on average (28.3 months),
- Wieand and the proposed rule have moderate durations (both around 32 months),
- No IA runs longest (36.5 months).

Interpretation: Scenario C shows that when early data are falsely optimistic, all rules risk continuing an ineffective treatment. The expected duration patterns reflect the aggressiveness of each futility boundary: O'Brien–Fleming stops earliest, while No IA never terminates early. The proposed rule behaves similarly to Wieand, indicating that its triggering mechanism does not improve robustness against misleadingly favourable early data.

6.3.4 Discussion

The scenarios examined above are intentionally extreme and not expected to occur frequently in practice. Their purpose is diagnostic: they show where the proposed rule can behave poorly and why. This helps clarify the limits of the approach rather than undermine its general usefulness.

Of the three cases, Scenario B is the most realistic and also the most concerning. When event rates are high, many events occur before patients reach 3 months of follow-up, so the two-thirds condition may never be met. As a result, the proposed rule can fail to trigger any futility analysis, even when the treatment is clearly ineffective or harmful. Because the main aim of futility monitoring is to stop such trials early, missing the opportunity to conduct an interim undermines the purpose of having one.

These results show that the behaviour of the rule depends strongly on accrual rate, baseline hazards, and the expected delay. Designers should therefore assess in advance whether the rule behaves like a standard futility rule, produces delayed interim analyses, or suppresses interim analyses altogether. This assessment can be conducted through routine simulation or by incorporating prior knowledge, for example about control survival or the expected treatment delay, into elicitation exercises.

Future work could consider alternative thresholds, different time cut-offs, or hybrid approaches that ensure at least one futility look while still accounting for delayed effects. Ultimately, the suitability of the proposed rule depends on the trial setting and the importance placed on avoiding premature futility decisions versus ensuring timely identification of lack of benefit.

Although the Korn–Freidlin modification offers a pragmatic safeguard against premature futility stopping, the investigation above shows that its behaviour depends sensitively on accrual patterns, the baseline event rate, and the shape of the delayed effect. In most realistic settings the rule behaves as intended, but its performance can degrade in edge cases where the proportion of post-delay events is difficult to predict or control. Importantly, the rule treats the timing of the delay as fixed and known, and therefore cannot adapt dynamically when the observed data suggest that the onset of benefit may occur earlier or later than anticipated.

These limitations highlight the need for a more flexible approach to interim monitoring—one that can incorporate uncertainty about both the magnitude and timing of the treatment effect and update its predictions as interim data accrue. Predictive probability provides exactly this capability. By integrating the elicited prior distributions for the control hazard, the delay duration, and the post-delay treatment effect with the observed interim dataset, PP yields a model-based assessment of the probability of eventual success under the delayed-effect structure. This motivates the transition from rule-based futility timing to predictive monitoring within an adaptive design framework.

6.4 Predictive Probability with Delayed Treatment Effects

As introduced in Section 5.3.2, Predictive Probability (PP) provides a model-based framework for interim decision-making, but its application requires explicit specification of the data-generating process for both the observed and future outcomes. In the presence of delayed treatment effects, this step becomes non-standard: the likelihood must accommodate a piecewise treatment hazard, and the predictive simulations must account for uncertainty in both the magnitude and timing of the delayed effect. This section develops the extension of PP to delayed-effect survival models.

At an interim analysis, the elicited priors for the control hazard, the post-delay treatment effect, and the delay time are updated using the observed data to obtain the joint posterior distribution. Posterior draws are then propagated forward to simulate the unobserved follow-up, construct the hypothetical final analysis, and evaluate the prespecified success criterion. Averaging these indicators yields the PP—the probability of ultimately achieving the primary endpoint conditional on the interim information.

To implement this procedure, we formulate explicit likelihoods for the delayed-effect models. As in Chapter 4, we consider exponential and Weibull control hazards with a piecewise treatment hazard to represent the delay. These likelihoods, combined with the elicited priors, permit posterior computation via MCMC and form the basis of the predictive probability calculations.

6.4.1 Likelihood

For patient i, let x_i denote the observed follow-up time, $y_i \in \{0, 1\}$ the event indicator, and $z_i \in \{0, 1\}$ the treatment indicator. Under a delayed treatment effect with delay duration τ, treated patients contribute either through the pre-delay or post-delay haz-

ard depending on whether $x_i \leq \tau$ or $x_i > \tau$. For censored observations ($y_i = 0$), only the cumulative hazard contributes.

Define the index sets

$$\mathcal{C} = \{i : z_i = 0\}, \qquad \mathcal{E}_{\leq\tau} = \{i : z_i = 1,\, x_i \leq \tau\}, \qquad \mathcal{E}_{>\tau} = \{i : z_i = 1,\, x_i > \tau\}.$$

6.4.1.1 Exponential Model

Under the exponential model (Section 4.4.1), the control hazard is $h_c(t) = \lambda_c$. After the delay, the treatment hazard is scaled by a constant hazard ratio,

$$h_e(t) = \lambda_c \,\mathrm{HR}^*, \qquad t > \tau.$$

For patients in $\mathcal{C} \cup \mathcal{E}_{\leq\tau}$, the hazard is λ_c throughout, giving

$$L_i = \lambda_c^{y_i} \exp(-\lambda_c x_i).$$

For patients in $\mathcal{E}_{>\tau}$, the cumulative hazard splits into a λ_c segment of length τ and a post-delay segment of length $x_i - \tau$:

$$H_i(x_i) = \lambda_c\tau + \lambda_c\mathrm{HR}^*\,(x_i - \tau),$$

and their likelihood contribution is

$$L_i = (\lambda_c\mathrm{HR}^*)^{y_i} \exp[-\lambda_c\tau - \lambda_c\mathrm{HR}^*(x_i - \tau)]\,.$$

Let $\boldsymbol{\theta} = (\lambda_c, \mathrm{HR}^*, \tau)$. The full likelihood is

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_{i\in\mathcal{C}\cup\mathcal{E}_{\leq\tau}} (\lambda_c^{y_i} \exp(-\lambda_c x_i)) \prod_{i\in\mathcal{E}_{>\tau}} \left[(\lambda_c\mathrm{HR}^*)^{y_i} \exp\big(-\lambda_c\tau - \lambda_c\mathrm{HR}^*(x_i - \tau)\big)\right].$$

For computation and optimisation, inference is performed using the log-likelihood, which converts products into sums and improves numerical stability. This is given in Appendix A.3.

6.4.1.2 Weibull Model

Under the Weibull specification (Chapter 4), the control-arm hazard, cumulative hazard, and survival functions are

$$h_c(t) = \gamma_c\lambda_c^{\gamma_c} t^{\gamma_c - 1}, \qquad H_c(t) = (\lambda_c t)^{\gamma_c}, \qquad S_c(t) = \exp\{-(\lambda_c t)^{\gamma_c}\}.$$

We adopt the common-shape assumption $\gamma_e = \gamma_c$, so that the post-delay treatment effect can be expressed as a constant hazard ratio:

$$h_e(t) = \mathrm{HR}^* \, h_c(t), \qquad t > \tau.$$

For all $i \in \mathcal{C} \cup \mathcal{E}_{\le\tau}$, the likelihood contribution is

$$L_i = \left(\gamma_c \lambda_c^{\gamma_c} x_i^{\gamma_c - 1}\right)^{y_i} \exp\left(-(\lambda_c x_i)^{\gamma_c}\right).$$

For $i \in \mathcal{E}_{>\tau}$, the cumulative hazard decomposes into a pre-delay segment plus a post-delay segment scaled by HR^*:

$$H_i(x_i) = H_c(\tau) + \mathrm{HR}^* \left[H_c(x_i) - H_c(\tau)\right] = (\lambda_c \tau)^{\gamma_c} + \mathrm{HR}^* \left((\lambda_c x_i)^{\gamma_c} - (\lambda_c \tau)^{\gamma_c}\right).$$

The likelihood contribution is therefore

$$L_i = \left(\mathrm{HR}^* \gamma_c \lambda_c^{\gamma_c} x_i^{\gamma_c - 1}\right)^{y_i} \exp\left[-(\lambda_c \tau)^{\gamma_c} - \mathrm{HR}^* \left((\lambda_c x_i)^{\gamma_c} - (\lambda_c \tau)^{\gamma_c}\right)\right].$$

Let $\boldsymbol{\theta} = (\lambda_c, \gamma_c, \mathrm{HR}^*, \tau)$ denote the parameter vector. The full likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}) = &\prod_{i \in \mathcal{C} \cup \mathcal{E}_{\le\tau}} \left[\left(\gamma_c \lambda_c^{\gamma_c} x_i^{\gamma_c - 1}\right)^{y_i} \exp\left(-(\lambda_c x_i)^{\gamma_c}\right)\right] \\ &\times \prod_{i \in \mathcal{E}_{>\tau}} \left[\left(\mathrm{HR}^* \gamma_c \lambda_c^{\gamma_c} x_i^{\gamma_c - 1}\right)^{y_i} \exp\left(-(\lambda_c \tau)^{\gamma_c} - \mathrm{HR}^* \left((\lambda_c x_i)^{\gamma_c} - (\lambda_c \tau)^{\gamma_c}\right)\right)\right]\end{aligned}$$

As with the exponential model, inference proceeds using the log-likelihood, given in Appendix A.3.

6.4.2 Posterior Updating Under Delayed Treatment Effects

Given the likelihood expressions defined above, Bayesian updating proceeds by combining the likelihood with prior distributions on all model parameters, including the delay time τ. Let $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid \mathcal{D})$ denote the prior and posterior densities for the parameter vector $\boldsymbol{\theta}$. At the interim analysis, the posterior is

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{int}}) \;\propto\; \mathcal{L}(\boldsymbol{\theta} \mid \mathcal{D}_{\text{int}}) \, p(\boldsymbol{\theta}),$$

where $\mathcal{D}_{\text{int}}$ denotes the event and censoring times observed up to the interim look.

Under the exponential delayed–effect model, the parameter vector is

$$\boldsymbol{\theta} = (\lambda_c, \mathrm{HR}^*, \tau).$$

Under the Weibull specification, the parameter vector additionally includes the shape parameter,

$$\boldsymbol{\theta} = (\lambda_c, \gamma_c, \mathrm{HR}^*, \tau).$$

Posterior inference is obtained by sampling from $p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{int}})$ using `rjags` (Plummer, 2025). The full JAGS model specifications for both the exponential and Weibull cases are provided in Appendix B.

At MCMC iteration m, the draw

$$\boldsymbol{\theta}^{(m)} = (\lambda_c^{(m)}, \gamma_c^{(m)}, \mathrm{HR}^{*(m)}, \tau^{(m)}) \quad \text{or} \quad (\lambda_c^{(m)}, \mathrm{HR}^{*(m)}, \tau^{(m)})$$

(for the Weibull and exponential models, respectively) fully determines the corresponding piecewise hazard and cumulative hazard functions for both treatment arms.

Convergence is assessed using standard diagnostics—trace plots, effective sample sizes, and the $\widehat{R}$ statistic. The resulting posterior sample,

$$\{\boldsymbol{\theta}^{(m)} : m = 1, \ldots, M\},$$

provides the basis for the predictive probability calculations described in the next section.

6.4.3 Calculating PP with Delayed Treatment Effects

Following posterior updating at the interim analysis, the goal is to quantify the probability that the final analysis will reject H_0, accounting for uncertainty in both the model parameters and all future unobserved outcomes. Let $\boldsymbol{\theta}$ denote the full parameter vector and $\tilde{\mathcal{D}}$ the unobserved survival and censoring outcomes between the interim and final analyses.

Unlike standard proportional-hazards settings, computing PP under delayed treatment effects is non-trivial. At interim, patients belong to distinct risk-set categories, each governed by different hazard functions depending on treatment assignment and accumulated follow-up relative to the delay time τ:

- patients not yet enrolled (both arms);
- censored control patients;
- censored treated patients in the pre-delay region ($x_i < \tau$);
- censored treated patients in the post-delay region ($x_i > \tau$);

- patients who have already experienced the event.

This heterogeneity in data-generating mechanisms is what makes predictive calculations under delayed effects substantially more complex than in models with proportional hazards. Conditional on a posterior draw $\boldsymbol{\theta}^{(m)}$, future event times are simulated for all patients at risk — including both those already enrolled and those yet to be recruited — using the appropriate piecewise hazard structure, with administrative censoring enforced by the trial calendar. Explicit simulation formulae for the exponential and Weibull cases are given in Appendix A.4. The full computational procedure is formalised in Algorithm 7.

Algorithm 7 Predictive probability futility rule simulation under delayed treatment effects

Inputs: Per-arm sample sizes n_c, n_e; data generating mechanism $\mathcal{G}$; maximum events E; information fraction for interim F_1; futility threshold PP_{min}; efficacy boundary b; final efficacy boundary b_{final}; recruitment model $\mathcal{R}$; censoring model $\mathcal{C}$; number of posterior draws M; number of replicates N
for $i = 1, \ldots, N$ **do**
    Generate control and treatment event times from $\mathcal{G}$
    Apply recruitment model $\mathcal{R}$ to obtain calendar entry times
    Determine calendar time $E_{T,1}$ at which $\lfloor F_1 E \rfloor$ events have occurred
    Apply censoring model $\mathcal{C}$ at $E_{T,1}$ to form interim dataset $\mathcal{D}_{\text{int},i}$
    Compute interim test statistic $Z_{i,1}$
    if $Z_{i,1} > b$ **then**
        Set $U_i = 1$ and **continue** ▷ Stop for efficacy
    else
        for $m = 1, \ldots, M$ **do**
            Draw $\boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{int},i})$ via MCMC
            Simulate future event times $\tilde{\mathcal{D}}^{(m)}$ from $p(\tilde{\mathcal{D}} \mid \boldsymbol{\theta}^{(m)})$, applying treatment-specific and delay-dependent hazards
            Form completed dataset $\mathcal{D}_{\text{int},i} \cup \tilde{\mathcal{D}}^{(m)}$ and perform final analysis A
            Set $\mathbb{I}^{(m)} = \mathbf{1}(\text{analysis successful})$
        end for
        Compute $\widehat{\text{PP}}_i = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}^{(m)}$
        if $\widehat{\text{PP}}_i < \text{PP}_{\text{min}}$ **then**
            Set $U_i = 0$ ▷ Stop for futility
        else
            Proceed to final analysis and set $U_i = \mathbf{1}(Z_{i,\text{final}} > b_{\text{final}})$
        end if
    end if
end for
Estimate probability of success, where R denotes the event of a successful analysis:

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

This Monte Carlo estimator converges almost surely to the true predictive probability as $M \to \infty$. In practice, M in the range 1,000–10,000 yields adequate numerical stability for interim decision-making.

6.5 Group Sequential Designs with Delayed Treatment Effects

As introduced in Section 5.4, group sequential designs provide a framework for interim decision-making based on accumulating event data. However, their standard formulation assumes proportional hazards, and the operating characteristics of interim monitoring rules are typically derived analytically or via simulation under this assumption. In the presence of delayed treatment effects, the data-generating mechanism is more complex and analytical evaluation is no longer tractable. This section develops the extension of the group sequential simulation framework to delayed-effect survival models.

A key challenge in applying group sequential methods to delayed-effect settings is the risk of premature futility stopping. Standard boundaries are calibrated under the assumption that the treatment effect is present from the outset, so that interim test statistics carry meaningful information about the eventual treatment benefit. When the treatment effect is delayed, however, early interim data are dominated by events from the pre-delay period, during which the treatment and control hazards are equal. Interim test statistics computed at this stage may be attenuated or may even favour the control arm (Li et al., 2021; Ghosh et al., 2021), and standard futility boundaries may therefore recommend stopping a trial that would ultimately be successful if allowed to continue (Korn and Freidlin, 2018; Wu et al., 2023).

A related complication concerns the interpretation of information fraction. In standard group sequential designs, information fraction is defined as the ratio of accumulated events to the planned maximum, and the log-rank statistic is known to follow an asymptotic normal distribution with variance proportional to the information fraction. Under delayed treatment effects, this relationship no longer holds exactly: early events carry less information about the treatment effect than later events, so equal increments in the number of events do not correspond to equal increments in statistical information (Wu et al., 2023; Li et al., 2021). This non-uniform information accrual further complicates the calibration of interim boundaries and reinforces the need for simulation-based evaluation of operating characteristics.

At each interim analysis, the observed data are used to compute a test statistic based on the accumulated events to date. This statistic is then compared against pre-specified efficacy and futility boundaries: if it exceeds the efficacy boundary the trial stops for success, if it falls below the futility boundary the trial stops for futility, and otherwise the trial continues to the next look. Under delayed treatment effects, the data generated between interim looks follow a piecewise hazard structure that depends on treatment assignment and accumulated follow-up relative to the delay time τ, requiring simulation to evaluate operating characteristics. The full computational procedure is formalised

in Algorithm 8.

Algorithm 8 Group sequential design simulation under delayed treatment effects

Inputs: Per-arm sample sizes n_c, n_e; priors $\pi(\boldsymbol{\theta_c})$, $\pi(\tau)$, $\pi(\text{HR}^*)$; separation probability $P_{\mathcal{S}}$; delayed-effect probability P_{DTE}; maximum events E; information fractions $F_1, \ldots, F_L$; futility boundaries $a_1, \ldots, a_L$; efficacy boundaries $b_1, \ldots, b_L$; recruitment model $\mathcal{R}$; censoring model $\mathcal{C}$; primary analysis A; number of replicates N
for $i = 1, \ldots, N$ **do**
    Sample $\boldsymbol{\theta}_{c,i} \sim \pi(\boldsymbol{\theta_c})$
    if $u \sim \text{Uniform}(0,1) \geq P_{\mathcal{S}}$ **then**
        Set $\tau_i = 0$ and $\text{HR}_i^* = 1$ ▷ No treatment effect
    else if $u \sim \text{Uniform}(0,1) < P_{\text{DTE}}$ **then**
        Sample $\tau_i \sim \pi(\tau)$ and $\text{HR}_i^* \sim \pi(\text{HR}^*)$ ▷ Delayed treatment effect
    else
        Set $\tau_i = 0$ and sample $\text{HR}_i^* \sim \pi(\text{HR}^*)$ ▷ Immediate treatment effect
    end if
    Generate control and treatment event times using $\boldsymbol{\theta}_{c,i}$, τ_i, HR_i^*
    Apply recruitment model $\mathcal{R}$ to obtain calendar entry times
    for $j = 1, \ldots, L$ **do**
        Determine calendar time $E_{T,j}$ at which $\lfloor F_j E \rfloor$ events have occurred
        Apply censoring model $\mathcal{C}$ at $E_{T,j}$
        Compute interim test statistic $Z_{i,j}$
        if $Z_{i,j} < a_j$ **then**
            Set $U_i = 0$ and **break** ▷ Stop for futility
        else if $Z_{i,j} > b_j$ **then**
            Set $U_i = 1$ and **break** ▷ Stop for efficacy
        end if
    end for
    if trial did not stop early **then**
        Set $U_i = \mathbf{1}(Z_{i,L} > b_L)$ ▷ Final analysis
    end if
end for
Estimate probability of success, where R denotes the event of a successful analysis:

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

6.6 Example

The methods developed in this chapter provide the foundation for constructing adaptive designs in settings where treatment effects may emerge only after an initial delay. To illustrate how these components operate in practice, we now apply the framework to the exponential example introduced in Section 4.7.1. This worked example demonstrates how elicited priors feed into the posterior and predictive modelling, how interim timing and futility thresholds may be calibrated, and how the resulting adaptive design behaves under both fixed-effect and prior-predictive evaluations.

We consider a two-arm randomised superiority trial with uniform recruitment over a 24-month accrual period, 1:1 allocation, and overall survival (OS) as the primary endpoint. The confirmatory analysis is a one-sided log-rank test at the 2.5% significance level. Under the fixed design, this corresponds to a single final analysis at the planned number of events. In the adaptive designs explored below, the same inferential threshold is incorporated into a group-sequential framework, with efficacy boundaries calibrated to preserve the overall type I error rate across interim and final analyses.

For completeness, we restate the elicited prior distributions introduced in Section 4.7.1. The control-arm hazard rate λ_c is modelled via a meta-analytic predictive prior, approximated by a Gamma(14.2, 181) distribution. The prior probability of the survival curves separating is $P_{\mathcal{S}} = 0.9$, and the prior probability that a delayed effect exists is $P_{\text{DTE}} = 0.8$. Conditional on a delay, the onset time follows

$$\tau \sim \text{Gamma}(7.29,\ 1.76),$$

and the post-delay hazard ratio is assigned

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

These priors induce uncertainty not only about the magnitude of the treatment effect but also about its timing, and therefore play a central role in shaping the posterior distribution at interim and the predictive probability of ultimate success. In the following subsections, we show how these elicited components are integrated into (i) posterior updating at the interim analysis, (ii) simulation-based calibration of the PP futility threshold, and (iii) full forward simulation to evaluate the design's operating characteristics.

6.6.1 No Interim Analysis

We begin by examining the fixed design with no interim monitoring. This serves as a baseline against which the impact of introducing adaptive decision rules can be assessed.

Under the assumed recruitment schedule and event dynamics, a total of 400 patients per arm are enrolled, and the final log-rank analysis is conducted once 650 events have accrued. Evaluated under the prior predictive distribution, the probability of ultimately rejecting the null hypothesis is approximately 80% (up to Monte Carlo error). Because no interim analyses are performed, the sample size remains fixed at 800 patients and the mean trial duration is approximately 42 months.

These operating characteristics provide a reference point for interpreting the adaptive designs that follow, allowing direct comparison of how interim futility monitoring affects power, expected sample size, and trial duration.

6.6.2 Adaptive Design Using Predictive Probability

Having established the baseline fixed design, we now consider an adaptive alternative in which a single interim futility analysis is based on the predictive probability (PP) of eventual success. Control of the one-sided Type I error rate at 2.5% is maintained through a group sequential efficacy boundary, yielding a hybrid Bayesian–frequentist monitoring strategy.

Construction of this adaptive design follows the three-step workflow described in Section 5.6.3:

1. **Selecting the timing of the PP interim analysis**, ensuring sufficient information to produce a stable predictive probability while retaining the potential for meaningful efficiency gains;

2. **Calibrating a PP-based futility threshold** that balances early stopping for unpromising scenarios against preservation of power when treatment effects are plausible; and

3. **Evaluating the complete design**, embedding both the PP futility rule and group sequential efficacy monitoring, across a range of clinically relevant data-generating mechanisms.

This framework ensures that futility decisions incorporate uncertainty about delayed treatment effects, while efficacy decisions retain the familiar frequentist error control required in confirmatory settings.

6.6.2.1 Step 1: Choosing the Timing of the PP Look

Candidate information fractions between 0.20 and 0.80 (in increments of 0.10) were evaluated following the procedure in Section 5.6.3.1. For each candidate timing, interim datasets were generated from the prior predictive distribution under the delayed–effect model, the posterior was updated, and the predictive probability (PP) of ultimate success was computed.

Figure 6.5 shows histograms of the PP values at selected information fractions. Early analyses (e.g. IF = 0.20) yield PP distributions concentrated near 0.5, reflecting the limited information available when few events have accrued. Mid-trial analyses (IF = 0.40–0.60) exhibit more U-shaped distributions with increasing mass near 0 and 1, indicating meaningful discriminatory ability. Very late looks (IF = 0.70–0.80) are highly informative but occur after most recruitment has completed, reducing their operational value.

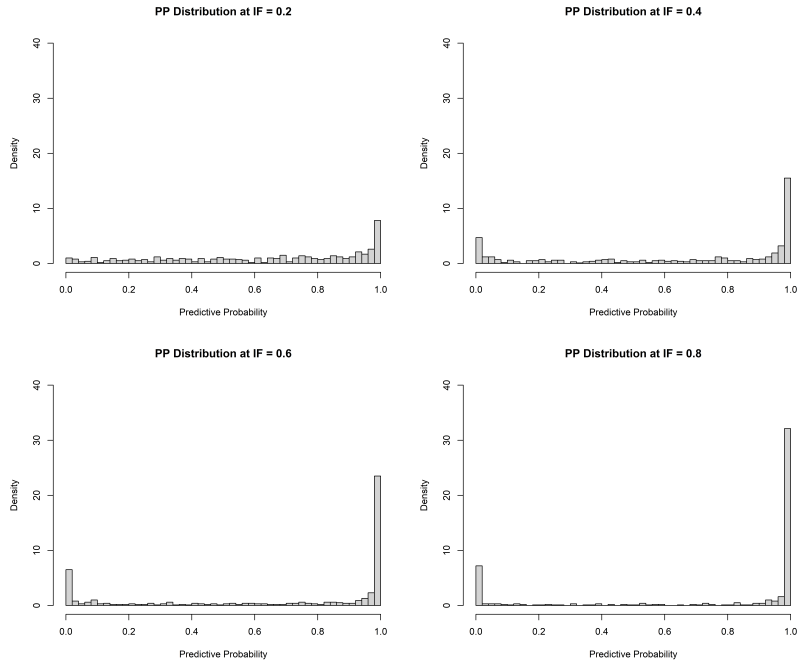


Figure 6.5: *Histograms of the predictive probability (PP) at selected information fractions (IF = 0.20, 0.40, 0.60, 0.80). Early looks provide weak discrimination, whereas mid-trial looks yield substantially more informative PP distributions.*

To quantify informativeness, we computed

$$P(\text{PP} < 0.10 \text{ or } \text{PP} > 0.90),$$

the probability that the interim analysis yields a near-certain futility or success decision. Table 6.3 summarises this metric and the corresponding expected calendar times.

Table 6.3: *Informativeness of the predictive probability (PP) at candidate information fractions. Informativeness increases sharply between IF = 0.30 and 0.60, after which gains are marginal and operational value decreases.*

Information Fraction (IF)	0.20	0.30	0.40	0.50	0.60	0.70	0.80
Informativeness	0.36	0.50	0.61	0.69	0.73	0.82	0.88
Interim Time (months)	12.2	15.5	18.6	21.3	24.2	27.2	30.9

Taken together, the results indicate that information fractions in the range 0.40–0.60 offer the best compromise between statistical discrimination and the opportunity for meaningful early stopping. We therefore select a single PP futility look at IF = 0.50.

6.6.2.2 Step 2: Calibration of the PP Futility Threshold

With the interim timing fixed at IF = 0.50, we calibrated the futility threshold c following the procedure in Section 5.6.3.2. To characterise the behaviour of the predictive probability (PP) under different treatment-effect assumptions, we simulated 2,000 interim datasets under three scenarios:

1. **Null scenario**: no treatment effect.
2. **Fixed-delay alternative**: 4-month delay, then post-delay HR = 0.6.
3. **Immediate-effect alternative**: proportional hazards with HR = 0.6.

For each dataset, posterior updating was performed and the PP of success at the final analysis was computed. Figure 6.6 displays the resulting PP distributions. Under the null, PP values are concentrated near zero; under both alternative scenarios, PP values cluster strongly near one, with the fixed-delay case exhibiting a slightly longer lower tail due to the attenuated early effect.

To inform threshold selection, Table 6.4 reports $\Pr(\text{PP} < c)$ for a range of candidate cutoffs. Under the null scenario, the probability of crossing a futility threshold rises sharply with c (e.g. 56% at $c = 0.05$, 78% at $c = 0.20$, 85% at $c = 0.30$). In contrast, both alternative scenarios show extremely low probabilities of incorrectly crossing the futility boundary: below 0.3% for the fixed-delay case and effectively zero for the immediate-effect case for $c \leq 0.25$.

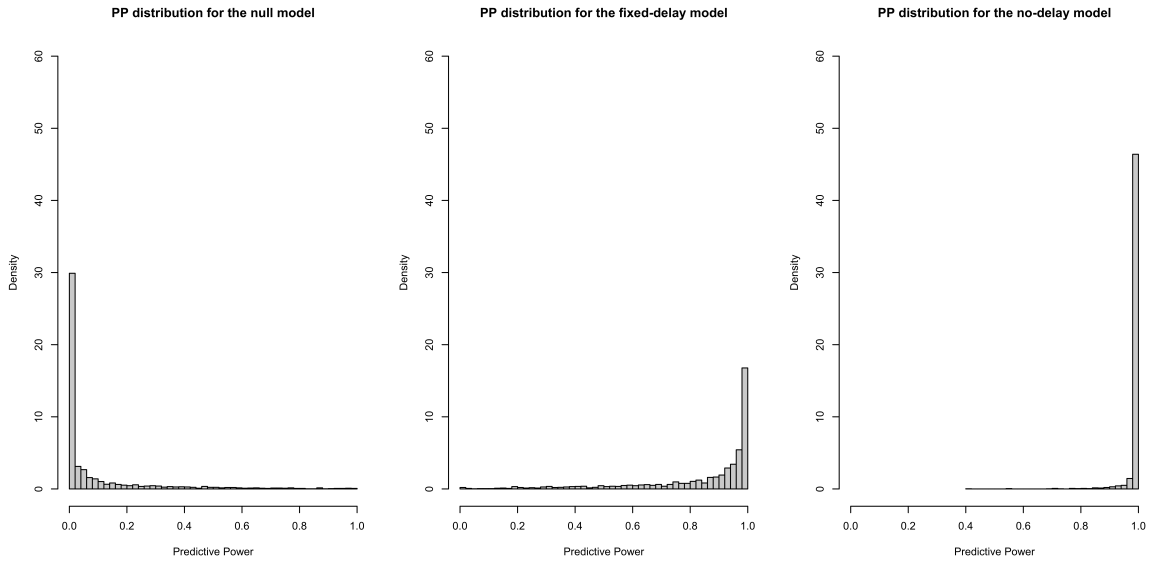


Figure 6.6: *Predictive probability (PP) at the interim analysis (IF = 0.50) under the null, fixed-delay, and immediate-effect scenarios. The null scenario yields PP values near zero, while both alternatives are concentrated near one, showing clear separation across clinically relevant settings.*

These results indicate that thresholds in the range $c \in [0.15, 0.25]$ effectively discriminate between null and alternative scenarios. We selected $c = 0.20$ as a conservative yet efficient choice: it stops approximately 78% of null trials while maintaining a false-futility rate below 0.3% for the fixed-delay alternative and 0% for the immediate-effect scenario.

Overall, the strong separation of PP distributions across scenarios highlights the suitability of predictive probability as a futility criterion, even when the treatment effect may be delayed.

6.6.2.3 Step 3: Evaluation of the Full Adaptive Design

With the interim timing fixed at IF = 0.50 and the futility threshold set to $c = 0.20$, we evaluated the full adaptive design under four data-generating scenarios:

1. **S1 (Null)**: no treatment effect.
2. **S2 (Immediate alternative)**: proportional hazards with HR = 0.6.
3. **S3 (Delayed alternative)**: 4-month delay followed by HR = 0.6.

Table 6.4: *Estimated* $\Pr(\text{PP} < c)$ *at the interim analysis (IF = 0.50) under each scenario, based on 2,000 simulations.*

Threshold c	Null	Fixed-delay	No-delay
0.05	0.5615	0.0005	0.0000
0.10	0.6665	0.0015	0.0000
0.15	0.7450	0.0025	0.0000
0.20	0.7780	0.0025	0.0000
0.25	0.8225	0.0045	0.0000
0.30	0.8495	0.0055	0.0000
0.35	0.8770	0.0085	0.0000
0.40	0.8860	0.0110	0.0005
0.45	0.9070	0.0155	0.0010
0.50	0.9235	0.0180	0.0010

4. **S4 (Prior predictive mixture)**: uncertainty integrated over the elicited prior distributions.

These scenarios span both fixed effect assumptions (S1–S3) and the uncertainty structure implied by the elicitation (S4). In practice, a comprehensive design exercise would explore a broader range of hazard ratios, delay distributions, and recruitment patterns; however, the four scenarios considered here capture the principal operating regimes relevant to delayed-onset survival studies. We compared four monitoring strategies:

- **D1: Fixed design** with no interim monitoring.
- **D2: Group sequential design (GSD)** with a single efficacy look at IF = 0.75. A one-sided $\alpha = 0.025$ spending function allocates 0.0125 at the interim look and 0.025 at the final analysis, giving efficacy boundaries

$$Z > 2.241 \quad \text{(interim)}, \qquad Z > 2.047 \quad \text{(final)}.$$

- **D3: Hybrid PP–GSD design**. Uses the same efficacy boundaries as D2, supplemented with the predictive-probability futility rule

$$\text{stop for futility at IF } = 0.50 \text{ if } PP < 0.20.$$

- **D4: α- and β-spending GSD**. The α-spending matches D2. A β-spending function allocates 0.05 at IF = 0.50 and 0.10 at the final analysis, implying a non-binding futility boundary

$$Z < 0.757 \quad \text{(interim)}.$$

The operating characteristics of each design were evaluated via Monte Carlo simulation using Algorithms 5, 8, and 7 for D1, D2/D4, and D3 respectively. For scenarios S1–S3, the prior sampling steps in each algorithm are replaced by fixed parameter values corresponding to the assumed data-generating mechanism.

For each design–scenario combination, we simulated 100,000 Monte Carlo replications to estimate:

- $P(\text{reject } H_0)$, which corresponds to type I error (under S1), power (under S2–S3) or assurance (under S4),
- probabilities of early stopping for efficacy and futility,
- expected sample size, and
- expected trial duration.

Table 6.5 summarises the resulting operating characteristics.

6.6.2.4 Results

The operating characteristics for all design–scenario combinations are presented in Table 6.5. Several clear patterns emerge.

6.6.2.4.1 Scenario S1

Both adaptive designs with futility monitoring (D3 and D4) deliver substantial efficiency gains relative to designs without futility (D1 and D2). Design D3 stops 84% of trials at the interim look and reduces the expected sample size to 686 participants, whereas D4 stops 79% with an expected sample size of 694. These reductions translate into markedly shorter expected durations (22–23 months). The small differences between D3 and D4 reflect the differing calibration principles: D3's PP-based rule is tuned to minimise false-futility under plausible alternatives, while D4's β-spending rule targets overall Type II error. The type I error remains close to nominal for all designs.

6.6.2.4.2 Scenario S2

Under a clinically meaningful delayed benefit, D1 and D2 achieve the highest power (0.991 and 0.988 respectively), with D2 also enabling early efficacy stopping in 81.6% of trials, substantially reducing expected duration from 40.6 to 30.3 months. D3 maintains high power (0.969) with a very low false-futility probability of 0.023, confirming that the predictive probability rule is appropriately conservative when the treatment effect

Table 6.5: *Operating characteristics of the four monitoring strategies (D1: no interim analysis; D2: GSD efficacy-only; D3: GSD with predictive probability futility; D4: GSD with β-spending futility) under the four data-generating scenarios S1–S4. Estimates are based on 100,000 simulated trials. P(Early Fut.) is not applicable for D2 as it does not include a futility stopping rule. ESS denotes expected sample size; duration is reported in months.*

Scenario	Design	P(Reject H_0)	P(Early Fut.)	P(Early Eff.)	ESS	Duration
S1: Null	D1	0.025	—	—	800.0	35.5
	D2	0.025	—	0.012	800.0	35.4
	D3	0.021	0.837	0.012	686.4	22.3
	D4	0.023	0.786	0.012	693.5	23.1
S2: 4-month delay, then HR = 0.6	D1	0.991	—	—	800.0	40.6
	D2	0.988	—	0.816	800.0	30.3
	D3	0.969	0.023	0.813	797.6	29.9
	D4	0.914	0.083	0.801	791.4	28.8
S3: HR = 0.6	D1	1.000	—	—	800.0	42.1
	D2	1.000	—	1.000	800.0	29.2
	D3	1.000	0.000	1.000	800.0	29.2
	D4	1.000	0.000	1.000	800.0	29.2
S4: Elicited Priors	D1	0.801	—	—	800.0	42.0
	D2	0.793	—	0.652	799.3	32.7
	D3	0.750	0.196	0.646	772.0	29.5
	D4	0.740	0.205	0.645	773.6	29.2

is present but delayed. By contrast, D4 is considerably more prone to false futility (0.083), reducing power to 0.914 despite similar early efficacy stopping behaviour. This highlights a key vulnerability of β-spending futility in delayed-effect settings: when early interim data do not yet reflect the true treatment benefit, the nominal Type II error allocation can be overly aggressive, leading to premature trial termination.

6.6.2.4.3 Scenario S3

When the treatment effect is strong and present from time zero, all designs achieve essentially full power. Designs D2–D4 produce nearly identical early stopping patterns and operational metrics, reflecting the fact that the signal is sufficiently pronounced for both predictive and spending-based futility rules to permit continuation. Unsurprisingly, efficiency gains relative to D1 arise entirely from early efficacy stopping rather than futility.

6.6.2.4.4 Scenario S4

When performance is averaged over the elicited uncertainty, both futility rules reduce the prior predictive probability of success relative to designs without futility monitoring (0.750 for D3 and 0.740 for D4, compared with 0.801 for D1 and 0.793 for D2).

This behaviour is expected: the prior assigns non-negligible mass to weak or moderately delayed effects, under which the predictive probability or β-spending rules may recommend stopping for futility. D3 yields an expected sample size of 772 and expected duration of 29.5 months, while D4 yields similar values of 774 and 29.2 months respectively.

A limitation of reporting only the prior predictive probability of success is that it does not distinguish between trials that correctly reject H_0 when the treatment is genuinely effective and those that incorrectly reject H_0 when it is not. To address this, we define a minimum clinically important difference (MCID) as a threshold on the post-delay hazard ratio HR*: a trial outcome is classified as a correct decision if the trial rejects H_0 when the sampled HR* < MCID (true positive) or fails to reject H_0 when the sampled HR* $\geq$ MCID (true negative). Since the true data-generating mechanism is known at each simulation replicate under S4, this classification is well-defined. Table 6.6 reports the proportion of correct decisions across a range of clinically plausible MCID thresholds.

Table 6.6: *Proportion of correct decisions under Scenario S4 (elicited priors) for each monitoring strategy across a range of MCID thresholds. A correct decision is defined as rejecting H_0 when the sampled HR* < MCID and failing to reject H_0 when HR* $\geq$ MCID.*

MCID	D1	D2	D3	D4
0.70	0.842	0.843	0.827	0.816
0.75	0.898	0.895	0.866	0.854
0.80	0.915	0.909	0.872	0.860
0.85	0.913	0.906	0.865	0.854
0.90	0.905	0.897	0.855	0.844

All designs achieve their highest correct decision rates at MCID values between 0.80 and 0.85, slightly above the designed hazard ratio of 0.75. Designs D1 and D2 perform similarly throughout, achieving a peak of approximately 91%, while D3 and D4 are somewhat lower, peaking at around 87% and 86% respectively. The reduction in correct decisions for D3 and D4 relative to D1 and D2 reflects the additional risk of false futility stopping when the true treatment effect is present but modest or delayed. Notably, D3 consistently outperforms D4 across all MCID thresholds, suggesting that the predictive probability futility rule makes marginally better decisions than the β-spending rule when evaluated against the elicited prior.

6.6.2.4.5 Summary

The simulation results confirm that the decisions made in Steps 1–2—namely, selecting IF = 0.50 for the futility look and calibrating the PP threshold to $c = 0.20$—produce

the intended operating characteristics: (i) substantial efficiency gains under the null, (ii) strong preservation of power under both delayed and immediate alternatives, and (iii) behaviour under the prior predictive mixture that faithfully reflects the distribution of plausible treatment effects. The comparison with D4 also underscores the advantage of a model-based PP rule in delayed-effect settings, where calendar-time spending rules may misrepresent the evidential content of the interim data.

6.6.3 Summary of the Example

This example illustrates how the three development steps of timing selection, threshold calibration, and evaluation across multiple scenarios combine to yield a coherent and operationally attractive adaptive monitoring strategy. Several broader themes emerge that help explain the practical value of predictive probability futility rules in time to event settings with possible non proportional hazards.

A central finding is the substantial efficiency gain under the null. Because the predictive distribution for success rapidly concentrates near zero when no treatment effect is present, most null trials terminate early, substantially reducing expected sample size and duration. These gains arise without compromising type I error control at the final analysis, reflecting the deliberate calibration of the PP threshold: conservative enough to avoid spurious futility stopping under modest early fluctuations, yet sufficiently aggressive to curtail unproductive trials when posterior evidence consistently disfavors benefit.

Equally important is the behaviour under delayed treatment effects. The design rarely stops effective treatments prematurely, even when early survival curves exhibit little separation. This robustness stems from the structure of the predictive calculation, which integrates uncertainty about both the magnitude *and* timing of the treatment effect and projects forward the anticipated accumulation of future events. As a result, the PP naturally accommodates delayed-onset benefit in a way that interim criteria based on proportional-hazards extrapolation cannot.

The comparison with the β-spending rule highlights the practical implications of these modelling differences. Although both approaches improve efficiency under the null, the conditional-power framework underlying the β-spending rule tends to overreact to early underperformance when effects are delayed, leading to avoidable loss of power. By implicitly assuming immediate proportional hazards, the β-spending criterion is vulnerable to early misestimation of the hazard ratio. The PP rule, in contrast, treats the onset of benefit as uncertain and therefore maintains power in delayed-effect scenarios.

Performance under model uncertainty, represented here through the elicited prior mixture, reinforces this conclusion. Both futility rules reduce the overall probability of

success relative to a non-adaptive design, as expected when prior mass is placed on weak or negligible effects. However, the PP-based design more faithfully reflects the prior beliefs regarding plausible benefit and delay patterns, yielding slightly higher prior predictive success. This alignment between the decision rule and the elicited distribution is conceptually attractive and practically meaningful, especially in early development programmes where decision-making under uncertainty is explicit.

Finally, the example demonstrates that Bayesian interim monitoring can be successfully embedded within a frequentist group-sequential framework while retaining the inferential guarantees required for regulatory review. Type I error is preserved through the GSD efficacy boundaries; the PP futility rule is transparent and model-based; and the joint behaviour of the hybrid design can be comprehensively quantified through simulation. Although only four representative scenarios are presented here, a full design exercise would typically examine a richer scenario grid, including alternative delay durations, attenuated post-delay effects, and key operational sensitivities, to ensure robustness across plausible clinical settings.

Taken together, these results show that PP-based futility monitoring can substantially enhance efficiency without compromising performance in complex time-to-event settings, making it a pragmatic and principled tool for modern oncology trial design.

6.6.4 Software Implementation

All computational procedures described in this chapter were implemented using the `DTEAssurance` R package developed as part of this research. The package provides a workflow that mirrors the three-step construction of the adaptive design:

- **Timing calibration.**
  The function `calibrate_PP_timing()` automates prior-predictive simulation of interim datasets across candidate information fractions. For each candidate timing, it updates the posterior under the delayed-effect model, evaluates the predictive probability (PP), and summarises the discriminatory behaviour of the PP distribution. This supports principled selection of an information fraction at which the futility analysis is both informative and operationally meaningful.

- **Threshold calibration.**
  The function `calibrate_PP_threshold()` evaluates the empirical distribution of the PP at the chosen interim time under multiple data-generating scenarios, including the null, fixed-delay alternatives, and immediate-effects models. These distributions guide selection of a futility threshold that balances protection against false-negative stopping with efficiency under unpromising scenarios.

- **Evaluation of the full adaptive design.**
  The function `calc_dte_assurance_adaptive()` implements the complete hybrid monitoring strategy, combining the calibrated PP futility rule with the specified group-sequential efficacy boundaries. It simulates full trial trajectories, applies the interim and final decision rules, and returns design-level operating characteristics including power, type I error, early stopping probabilities, expected sample size, and expected trial duration.

The package is accompanied by an interactive `Shiny` application, which exposes the same computational workflow through a graphical interface. The app enables users to specify elicited priors, explore different delay structures, examine PP timing and threshold behaviour, and visualise operating characteristics for both fixed and adaptive designs. This facilitates communication with clinicians and trial statisticians, and provides a reproducible, accessible platform for practical design exploration.

6.7 Discussion

This chapter has brought together elicited prior distributions, delayed–effect survival models, and adaptive monitoring strategies into a framework for designing time-to-event trials when treatment effects may emerge only after an initial delay. Building on the assurance methodology of Chapters 2 and 4 and the adaptive tools of Chapter 5, the work here demonstrates how predictive probability and group sequential methods can be jointly deployed to address the inferential and operational challenges that delayed treatment effects introduce.

A first contribution is the diagnostic assessment of the Korn–Freidlin timing rule for futility monitoring. While attractive for its pragmatic attempt to avoid premature stopping in the presence of delays, our exploration shows that its behaviour can depend sensitively on accrual patterns, baseline hazards, and the assumed onset time of treatment benefit. Although the rule works well in many settings, it can underperform or fail to trigger when early events dominate, indicating that timing restrictions alone are insufficient when uncertainty about delay is substantial. This motivates the need for more flexible, model-based interim criteria.

The second contribution is the development of a predictive-probability futility rule that explicitly incorporates elicited uncertainty regarding control survival, delay duration, and post-delay treatment effect. By updating the posterior at the interim and simulating forward the unobserved event process, the PP criterion directly targets the probability of ultimate trial success—even when most accumulated information lies in the pre-delay region. When coupled with group sequential efficacy monitoring, this hybrid Bayesian–frequentist framework preserves strict type I error control while enabling

early stopping decisions that remain coherent under delayed or evolving treatment effects.

A third contribution of the chapter is the software implementation accompanying the methodology. The `DTEAssurance` R package encapsulates the full workflow, including predictive probability timing calibration, threshold selection, and evaluation of hybrid adaptive designs, and provides a reproducible platform for design exploration. The associated `Shiny` application enables practitioners to visualise delay structures, interrogate predictive behaviour, and communicate design choices to clinical collaborators. Together, these tools translate the methodological developments into a form suitable for practical trial design and regulatory discussion, thereby lowering the barrier to adoption in applied settings.

The worked example illustrates how these components interact in practice, showing how predictive probability timing influences discriminative power, how the threshold balances efficiency under the null with protection against false negative stopping, and how Bayesian and frequentist decision rules jointly determine the operating characteristics of the overall design. Although the illustration is limited to a small set of scenarios for clarity, the workflow is intended to support broader scenario assessment, including alternative delay models, weaker post delay effects, and operational uncertainties, to ensure robustness in realistic settings.

Taken together, the methodological and software contributions of this chapter provide a coherent, extensible, and practically implementable framework for adaptive trial design in the presence of delayed treatment effects. The next chapter applies these tools to seven real oncology trials exhibiting non-proportional hazards, demonstrating how the framework can be used both retrospectively to understand trial performance and prospectively to guide future designs.

Chapter 7

Case Studies – Delayed Treatment Effects

This chapter is based on research using data from data contributors Roche and Bristol Myers Squibb that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.

7.1 Introduction

The preceding chapters have developed a methodology for designing, monitoring, and analysing survival trials in the presence of delayed treatment effects (DTEs). This chapter complements that development by examining empirical evidence from completed clinical trials to assess how the proposed concepts operate in practice. The aims are twofold: (i) to illustrate how delayed effects manifest in observed survival data, and (ii) to explore the potential implications of neglecting such effects when designing and conducting interim monitoring for confirmatory oncology trials.

Anonymised individual participant data (IPD) were obtained for seven completed phase III survival trials through the *Vivli* data-sharing platform (Vivli Project ID: 00009422; full details in Appendix C). These trials were selected based on published Kaplan–Meier curves exhibiting visual patterns consistent with delayed separation between treatment arms; as such, the sample is intentionally enriched for settings in which DTEs may arise.

Access to the Vivli datasets occurred late in the research programme. The data-use agreement was signed on 19 September 2024, with the first dataset (Roche) uploaded on 15 October 2024. Three Bristol Myers Squibb (BMS) studies became available on 7

January 2025, followed by four additional BMS studies on 30 March 2025 after follow-up correspondence. Documentation for the BMS datasets varied in completeness across studies. Because these materials arrived after the methodological development and all simulation work had been completed, the analyses in this chapter are descriptive and exploratory. Their purpose is to illustrate empirically that delayed treatment effects:

- arise in multiple contemporary oncology trials;
- can produce notable departures from the proportional hazards assumption;
- may increase the risk of premature or erroneous futility decisions at interim analyses; and
- can meaningfully reduce statistical power if unaccounted for at the design stage.

Individual participant data on overall survival (OS) were available for all seven trials, while progression free survival (PFS) data were available for five. A consistent analytical workflow was applied to each available endpoint, comprising visual diagnostics such as Kaplan–Meier inspection and Schoenfeld residual patterns, flexible estimation of time varying effects, and simulation based evaluation of operating characteristics.

One trial (denoted Trial G) is presented in detail to demonstrate the complete process from diagnostic assessment through to evaluation of operating characteristics under the proposed decision framework. This study exhibits a prototypical delayed effect pattern, with initial overlap between treatment groups followed by gradual divergence favouring the experimental arm. The remaining trials are summarised collectively to explore the extent to which similar behaviour arises across studies, with supplementary plots and outputs provided in Appendix C.

Taken together, these case studies offer empirical illustrations that support the methodological developments introduced in this thesis and emphasise the importance of accounting for potential delays in treatment effect when designing and monitoring survival trials.

7.2 Aims of the Chapter

This chapter applies a consistent analytical workflow to data from seven completed phase III survival trials. The analyses are designed to examine four key questions that align with the overarching objectives of this thesis:

1. **Is there empirical evidence of a delayed treatment effect?** We assess this using Kaplan–Meier curves, complementary log–log plots, and time-varying

hazard ratio estimates. These diagnostics are used to evaluate whether hazards are approximately proportional or whether a delay period precedes the emergence of treatment benefit.

2. **Do delayed-effect parametric models provide a superior description of the data compared with proportional hazards alternatives?** Model fit is evaluated using exponential and Weibull models, both with and without an explicit delay-time parameter. Adequacy is judged based on graphical agreement and information criteria.

3. **How might interim monitoring decisions be influenced by a delayed effect?** Retrospective group-sequential analyses are conducted to explore whether conventional futility monitoring rules could have led to premature termination in trials where benefit emerges only after a delay.

4. **What loss of statistical power arises if delayed effects are ignored at the design stage?** We quantify the discrepancy between proportional-hazards design assumptions and the realised survival behaviour, estimating the reduction in power attributable to non-proportional hazards.

Taken together, these four questions evaluate the robustness of conventional trial design and analysis methods in the presence of delayed treatment effects and assess whether explicit modelling or adaptive planning is required to mitigate associated operational and inferential risks.

7.3 Overview of Included Clinical Trials

Individual participant data (IPD) were obtained for seven completed phase III oncology trials through the *Vivli* data-sharing platform. A full record of associated scientific publications, protocols, and ethical approvals is provided in Appendix C. Each study evaluated an innovative systemic therapy against standard treatment, with time-to-event endpoints including overall survival (OS), progression-free survival (PFS), or both. Table 7.1 summarises the key characteristics of the included studies and the observed delay patterns.

All datasets were de-identified prior to release using the PhUSE date-shifting algorithm PHUSE, 2020. This procedure applies a participant-specific temporal offset, derived from each individual's earliest recorded timestamp, thereby preserving the relative ordering of assessments and events while removing absolute calendar time. Consequently:

- event-based analyses (e.g., survival curve estimation or hazard modeling) remain fully reproducible;

Table 7.1: *Key characteristics of the included studies and observed delay patterns. Delays are approximate based on visual assessment of Kaplan–Meier curves.*

Identifier	Number	Endpoint(s)	N	Events (OS, PFS)	OS Delay	PFS Delay
A	NCT02409342	OS	554	426, –	6.0	–
B	NCT01668784	OS, PFS	821	673, 745	0.0	15.0
C	NCT01721772	OS, PFS	418	295, 341	3.0	0.0
D[a]	NCT01844505	OS	629	390, –	3.0	–
			631	412, –	3.0	–
E	NCT01642004	OS, PFS	272	256, 219	0.0	2.5
F	NCT02477826		Excluded — event information unavailable			
G	NCT02105636	OS, PFS	361	339, 307	3.5	5.0
H	NCT01673867	OS, PFS	582	540, 506	7.0	8.0

[a] Trial D included two treatment regimens sharing a common control group.
These are analysed as separate pairwise comparisons (D1 and D2).

- accrual over calendar time, patterns of dropout, and the timing of original interim looks cannot be reconstructed;
- the *information fraction* must therefore be defined solely by the cumulative number of observed events.

These constraints limit the ability to reproduce the original monitoring schedule but remain compatible with the aims of this chapter, which concern the characterisation of delayed treatment effects and their implications for design and interim decision-making.

Visual inspection of the Kaplan-Meier curves confirms that several studies display features consistent with non-proportional hazards, including delayed separation of survival curves, curve crossing, and time-varying hazard ratios (full graphical outputs in Appendix C). In summary:

- delays in OS were observed in Trials A, C, D, G, and H;
- delays in PFS were evident in Trials B, E, G, and H.

The magnitude of delay varied considerably across trials, from negligible to more than six months, reflecting heterogeneity in disease mechanisms, treatment dynamics, and mechanisms of early non-responsiveness.

To illustrate the full analytical workflow in depth, Trial G is now examined as a representative case study before turning to the aggregate findings across all included trials.

7.4 Trial G

Trial G was selected to illustrate the complete analytical workflow because it demonstrates features commonly associated with delayed treatment effects, including early overlap of survival curves followed by progressive long-term separation. The trial enrolled 361 participants in a 2:1 allocation and evaluated overall survival (OS) and progression-free survival (PFS) as key time-to-event endpoints. Median follow-up in the anonymised dataset was 78 months, with 339 OS events and 307 PFS events observed.

7.4.1 Question 1: Evidence of a Delayed Treatment Effect

To determine whether a delayed onset of benefit was present in OS, four complementary diagnostics were applied:

(a) visual inspection of Kaplan–Meier (KM) curves;

(b) departing behaviour on the complementary log–log (clog–log) scale;

(c) time-varying hazard ratio trajectories over the event count; and

(d) a formal statistical test of the proportional hazards assumption.

7.4.1.1 Kaplan–Meier survival curves

Figure 7.1 shows the KM estimates for OS. The survival curves remain closely aligned for approximately the first 3–4 months before diverging in favour of the experimental arm. This early equivalence followed by subsequent improvement is characteristic of delayed treatment response.

7.4.1.2 Complementary log–log transformation

The clog–log plot in Figure 7.2 demonstrates a visible crossing of the transformed survival curves at approximately 4 months. Under proportional hazards, such curves would be approximately parallel. The observed crossing suggests non-proportionality and indicates different relative hazards in early and late phases of follow-up.

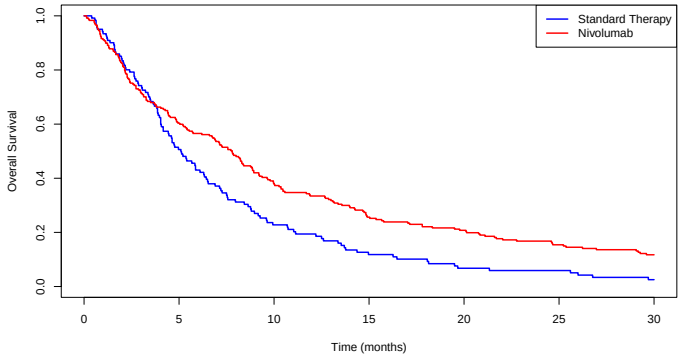


Figure 7.1: *Kaplan–Meier curves for overall survival in Trial G, illustrating delayed separation of treatment arms at approximately four months.*

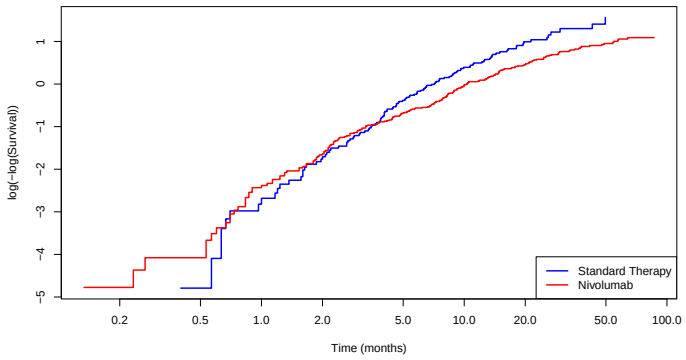


Figure 7.2: *Complementary log–log plot for overall survival in Trial G. The crossing of curves at approximately four months indicates non-proportional hazards.*

7.4.1.3 Time-varying hazard ratio

Figure 7.3 displays the instantaneous hazard ratio at each event time plotted against the cumulative number of events. The hazard ratio exceeds one during the early period, indicating worse short-term outcomes for the experimental arm, before decreasing below one at around 120 events (approximately one-third of total information). The subsequent monotonic decline reinforces evidence of a treatment benefit emerging only after a delay phase.

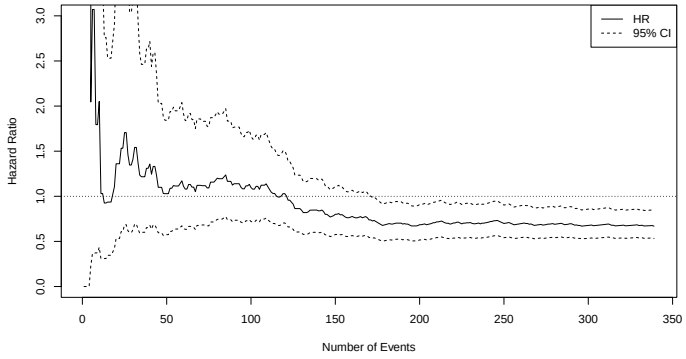


Figure 7.3: *Estimated hazard ratio plotted against cumulative number of overall survival events in Trial G. Initially, the hazard ratio is larger than one, and then crosses one at approximately 120 events, consistent with a delayed treatment effect.*

7.4.1.4 Proportional hazards test

To supplement graphical checks, the proportional hazards (PH) assumption was formally assessed using the global Grambsch–Therneau test based on Schoenfeld residuals. This test evaluates whether residuals show a systematic trend with time; under PH they should be time-independent. A small p-value indicates evidence against proportional hazards.

Table 7.2 reports the PH test results for Trial G:

Table 7.2: *Global Grambsch–Therneau test for proportional hazards in Trial G.*

Endpoint	Test statistic	p-value
OS	Schoenfeld residual trend	0.062
PFS	Schoenfeld residual trend	0.24

For OS, the p-value is borderline, consistent with the graphical evidence of a delayed effect but reflecting limited power during the early follow-up period where hazards appear similar. The PFS test does not indicate strong departure from proportional hazards at the global level; however, graphical diagnostics (Appendix C) reveal late separation suggestive of local non-proportionality. The Schoenfeld residual plot did not exhibit a constant trend over time, supporting the presence of mild non-proportionality in OS; a full diagnostic figure is provided in Appendix C.

7.4.1.5 Interpretation

All diagnostic tools indicate non-proportional hazards in OS for Trial G, with a meaningful delay period prior to treatment benefit. The PH test supports this conclusion but with reduced sensitivity due to early hazard equivalence. PFS shows a similar overall pattern, though with a longer delay phase (see Appendix C).

Taken together, these results motivate the use of modelling approaches that allow for delayed onset of effect and justify using Trial G as a representative case study for evaluating the design and monitoring implications of delayed treatment effects.

7.4.2 Question 2: Model Fit

To evaluate whether a piecewise specification (as in Chapter 4) provides an appropriate representation of the observed data, we fitted four parametric survival models:

- exponential proportional hazards;
- Weibull proportional hazards;
- piecewise exponential with piecewise-constant hazard ratio;
- piecewise Weibull with piecewise-constant hazard ratio.

All models were estimated by maximum likelihood using `flexsurv` (Jackson, 2016) in R.

7.4.2.1 Delay-time estimation

The delay parameter, τ, was estimated by a grid search that minimised the Akaike Information Criterion (AIC). Candidate values ranged from 2 to 10 months in 0.1-month increments. As shown in Figure 7.4, the AIC achieves its minimum at approximately $\tau = 3.3$ months for both piecewise distributions, in agreement with the graphical diagnostics in Section 7.4.1.

This estimate was therefore carried forward into all subsequent modelling and simulation analyses.

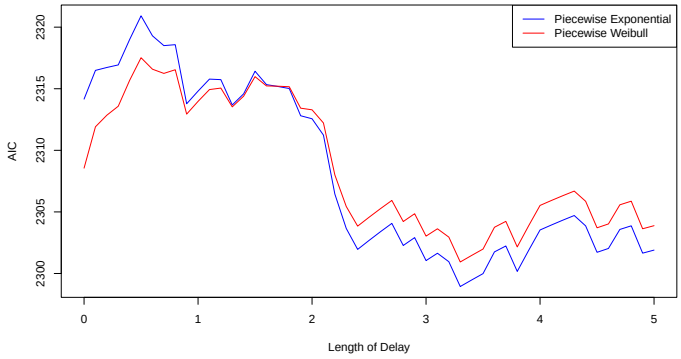


Figure 7.4: *Akaike Information Criterion (AIC) values across candidate delay durations for piecewise exponential and piecewise Weibull models (grid from 2 to 10 months in 0.1-month increments). Both models achieve the minimum AIC at approximately $\tau = 3.3$ months, confirming this as the best-supported estimate of the delay time and justifying its use in subsequent analyses.*

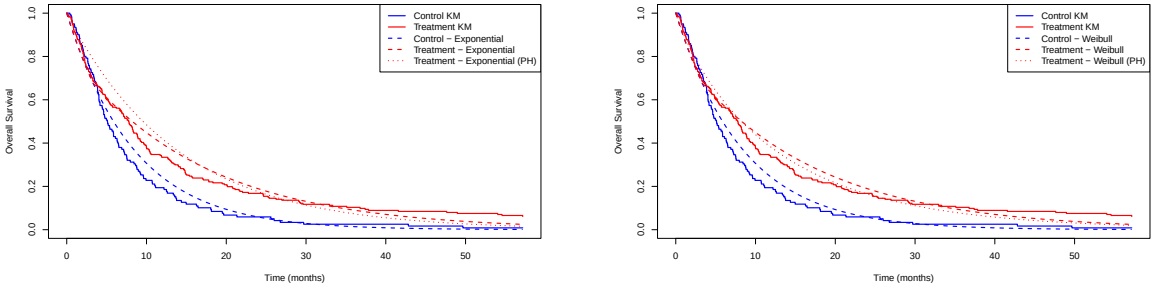
7.4.2.2 Model comparisons

Table 7.3 summarises parameter estimates and fit statistics. Both piecewise models provide improved fit relative to their PH counterparts, with the piecewise exponential model achieving the lowest AIC.

Table 7.3: *Parameter estimates and model fit for overall survival in Trial G. τ is the estimated delay time (months); λ_c and γ_c are the control scale and shape parameters; HR* is the post-delay hazard ratio.*

Trial	Model	τ	λ_c	γ_c	HR*	AIC
G	Exponential PH	–	0.1234	–	0.605	2273.7
	Weibull PH	–	0.1280	0.908	0.634	2270.2
	Piecewise exponential	3.3	0.1182	–	0.539	**2260.2**
	Piecewise Weibull	3.3	0.1187	1.019	0.522	2262.1

Figures 7.5a and 7.5b overlay the fitted piecewise and proportional hazards models against the Kaplan–Meier curves. The delayed separation is clearly better captured by the piecewise model, with the proportional hazards model failing to adequately describe the early period of no treatment effect.



(a) *Piecewise exponential fit.* **(b)** *Piecewise Weibull fit.*

Figure 7.5: *Observed Kaplan–Meier curves for overall survival in Trial G with fitted parametric models. The left panel shows the piecewise exponential fit and the right panel shows the piecewise Weibull fit, both allowing for a delay in treatment effect. In each panel, the corresponding proportional hazards model is also shown for comparison*

.

7.4.2.3 Interpretation

The estimated post-delay hazard ratio, $\text{HR}^* \approx 0.54$, indicates a meaningful long-term survival advantage once benefit emerges. In contrast, proportional hazards models average over the early and late phases, leading to weaker estimate magnitudes and inferior fit.

Overall, Trial G demonstrates that piecewise models can markedly improve inference when hazards are non-proportional and treatment benefit is delayed.

7.4.3 Question 3: Interim Decision Analysis

We next investigate how a delayed onset of treatment benefit might have influenced interim monitoring decisions under a standard group-sequential design. Because anonymisation prevents recovery of the original calendar-time looks, the monitoring process is indexed by the cumulative number of observed OS events. The *information fraction* is therefore defined as the proportion of the 339 events available in the de-identified dataset.

At each event time, we computed the one-sided log-rank Z-statistic based on all data accrued up to that point. To evaluate how the observed Z-trajectory aligns with typical group-sequential behaviour, we generated a set of hypothetical stopping boundaries under several plausible designs. These boundaries are not intended to reproduce the trial's actual monitoring plan; instead, they provide reference thresholds against which

the observed test statistic can be compared.

Specifically, we considered:

- a one-sided O'Brien–Fleming alpha-spending function ($\alpha = 0.025$) for a design with one interim and one final efficacy look; and
- Kim–DeMets beta-spending functions ($\beta = 0.10$) for futility, using three values of $\gamma \in \{0.75, 2.5, 6\}$, to span increasingly conservative rules.

Each spending function yields a distinct stopping boundary. All resulting efficacy and futility curves were plotted together and compared with the observed Z-process evaluated at every event time. This allows us to determine whether the trial would have crossed any of these reference boundaries under the corresponding hypothetical designs.

Table 7.4 reports the earliest boundary crossings (futility treated as non-binding). Figure 7.6 shows the observed Z-trajectory with the full set of hypothetical boundaries overlaid.

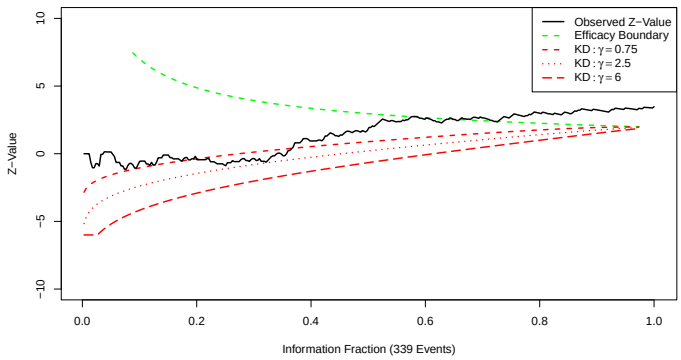


Figure 7.6: *O'Brien–Fleming alpha- (efficacy, green dashed) and Kim–DeMets beta- (futility, red dashed) spending boundaries evaluated at each information fraction, plotted against the cumulative number of events. The observed one-sided log-rank Z from Trial G (solid black) is overlaid. The Z-curve lies close to the futility region at low information and exceeds the efficacy boundary at 196 events (57.8%), while an aggressive futility rule ($\gamma = 0.75$) would have triggered stopping at 59 events (17.4%), illustrating the risk of premature futility stopping under delayed effects.*

Table 7.4: *Interim stopping behaviour for Trial G under one-sided O'Brien–Fleming ($\alpha = 0.025$) efficacy spending and Kim–DeMets ($\beta = 0.10$) futility spending. First crossing reported for each rule; futility treated as non-binding in this diagnostic analysis.*

Trial	Futility spending	Futility crossed	Efficacy crossed	Interpretation
G	$\gamma = 0.75$	59 (17.4%)	196 (57.8%)	Premature futility stop (benefit later emerged)
	$\gamma = 2.5$	–	196 (57.8%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	196 (57.8%)	Efficacy stop consistent with final outcome

7.4.3.1 Results and interpretation

Under aggressive futility spending ($\gamma = 0.75$), the futility boundary is crossed at 59 events (17.4% information), well before treatment benefit becomes apparent. This represents a *premature futility stop*—the trial would likely have been terminated early despite later evidence of benefit.

Under more conservative futility spending ($\gamma = 2.5$ or 6), no futility boundary is crossed. The observed Z later exceeds the efficacy boundary at 196 events (57.8% information), corresponding to an *efficacy stop consistent with the final outcome.*

These findings highlight a key operational risk in the presence of a delay period: early data are dominated by the pre-benefit phase, leading to inflated risk of falsely concluding lack of efficacy. When delayed effects are plausible, futility rules must be calibrated to avoid overweighting early hazard information.

7.4.4 Question 4: Power Loss

Finally, we investigated how the delayed treatment effect observed in Trial G impacted the empirical power of the study under its original design assumptions. The trial was designed to achieve 90% power (one-sided $\alpha = 0.025$) using a log-rank test under a proportional hazards assumption with a target hazard ratio of 0.67 (Ferris et al., 2016). The design specified a final analysis at 281 OS events with uniform recruitment over a 14-month period.

To quantify the effect of mis-specification, we evaluated three scenarios:

i) the proportional hazards design assumption (hazard ratio 0.67; no delay);

ii) the best-fitting delayed-effect model estimated from the observed data ($\tau = 3.3$ months; $\text{HR}^* \approx 0.54$);

iii) the assumed design hazard ratio (0.67) combined with the observed delay-time ($\tau = 3.3$ months).

All simulations used: 10,000 repetitions; 2:1 randomisation; uniform accrual over 14.2 months; and event-driven analysis at 281 OS events, consistent with the original design.

The resulting empirical power estimates are presented in Table 7.5.

Table 7.5: *Estimated power for Trial G under three scenarios: (i) the original design assumptions (no delay, HR=0.67), (ii) the best-fitting delayed-effect model from observed data (delay 3.3 months, HR*=0.54), and (iii) the design HR with the observed delay applied. A total of 281 OS events were required for the planned final analysis. All operating characteristics were estimated using $N = 10,000$ simulation replicates.*

	Design Parameters						Operating Characteristics	
Trial	**Scenario**	**Delay**	$\boldsymbol{\lambda_c}$	$\boldsymbol{\gamma_c}$	HR*	**Recruit.**	**Power (%)**	**Duration**
G	*(i)*	0.0	0.1155	–	0.67	14.2	90	25
	(ii)	3.3	0.1182	–	0.54	14.2	84	26
	(iii)	3.3	0.1155	–	0.67	14.2	49	24

7.4.4.1 Results and Interpretation

Under the proportional hazards (PH) assumptions used in the original design, the empirical power was approximately 90%, consistent with the planned operating characteristics.

When the observed delay-time and post-delay hazard ratio incorporated but the true post-delay hazard ratio was used, the power reduced modestly to around 84%.

In contrast, when the observed delay and assumed treatment effect was used, the power dropped sharply to approximately 49%.

The third scenario provides the clearest illustration of the consequences of PH misspecification: had the assumed effect size been correct but delayed, the design would have achieved only about half of its nominal power under a standard log-rank analysis. This highlights the sensitivity of conventional designs to the onset timing of treatment benefit, even when the ultimate magnitude of effect is unchanged.

These findings underscore two main points:

(a) **Delay-time primarily affects the information timing.** A substantial proportion of the planned events occur during the pre-benefit period, when the treatment and control hazards are similar. This early accumulation of events dilutes the overall test statistic and reduces effective information for detecting treatment benefit.

(b) **Trial G retained acceptable power because the realised post-delay benefit was stronger than assumed.** The favourable value of HR^* compensated for the inefficiency introduced by the delay, partially restoring power to near its target level.

From a design perspective, reliance on proportional hazards assumptions obscures uncertainty in the timing of treatment benefit, leading to overoptimistic expectations of statistical efficiency. This case demonstrates that unrecognised delays can markedly degrade power and threaten confirmatory success unless the design explicitly accommodates potential delays—either through increased sample size, delayed analyses, or piecewise modelling approaches.

7.5 Summary Across All Trials

The analytical workflow applied to Trial G was repeated for all studies listed in Table 7.1. The following subsections summarise the evidence obtained for each of the four key questions introduced in Section 2, with supporting figures and full numerical outputs reported in Appendix C.

7.5.1 Question 1: Evidence of Delayed Treatment Effects

Evidence for delayed separation of survival curves was evaluated using the same graphical and time-varying hazard techniques applied in Section 7.4.1. Visual inspection combined with non-proportional hazards indicators confirms that delayed effects are present in the majority of trials.

- **Clear evidence in OS**: Trials A, C, D1, D2, G, and H
- **Clear evidence in PFS**: Trials B, E, G, and H

Delay durations ranged from approximately 2 to 8 months depending on endpoint and disease context. Where no delay was detected, survival curves either separated

immediately or showed minimal divergence over follow-up. Global Grambsch–Therneau tests (Table 7.6) were consistent with these findings, providing further indication of hazard non-proportionality in several settings.

Table 7.6: *Global Grambsch–Therneau tests for proportional hazards by trial and endpoint (OS/PFS). A dash (–) indicates that the endpoint was not evaluated in that trial.*

Trial	Endpoint	p-value
A	OS	0.68
A	PFS	–
B	OS	0.25
	PFS	**0.00033**
C	OS	0.52
	PFS	**0.044**
D1	OS	**0.0010**
D1	PFS	–
D2	OS	0.066
D2	PFS	–
E	OS	0.76
	PFS	**0.0018**
G	OS	0.062
	PFS	0.24
H	OS	**0.00020**
	PFS	$\mathbf{2.96 \times 10^{-9}}$

7.5.2 Question 2: Model Fit

Piecewise exponential and Weibull models incorporating a delay parameter were compared against proportional hazards models in all trials. Table 7.7 summarises the results for OS; the corresponding PFS fits are provided in Table C.2.

Across the dataset:

- **Weibull models** provided the best fit in six of the eight trials, indicating the presence of time-varying baseline hazard shapes.
- **Models including a delay-time parameter** were preferred whenever a delayed effect was visually apparent.

Table 7.7: *Parameter estimates for overall survival (OS) across all trials under exponential and Weibull modelling frameworks. τ denotes the estimated delay time (months); λ_c and γ_c are the scale and shape parameters for the control arm; HR* is the post-delay hazard ratio from the best-fitting model. The lowest Akaike Information Criterion (AIC) within each trial is shown in **bold**.*

Trial	Model	τ	$\boldsymbol{\lambda_c}$	$\boldsymbol{\gamma_c}$	HR*	AIC
A	Exponential	–	0.0418	–	0.833	3559.5
	Weibull	–	0.0422	0.892	0.843	3553.6
	Piecewise exponential	5.3	0.0428	–	0.735	3553.8
	Piecewise Weibull	5.3	0.0420	0.923	0.794	**3552.3**
B	Exponential	–	0.0303	–	0.743	**6128.5**
	Weibull	–	0.0350	0.961	0.747	6128.9
	Piecewise exponential	0.0	0.0303	–	0.743	**6128.5**
	Piecewise Weibull	0.0	0.0305	0.961	0.747	6128.9
C	Exponential	–	0.0400	–	0.434	2651.2
	Weibull	–	0.0419	0.784	0.475	**2626.1**
	Piecewise exponential	2.0	0.0399	–	0.402	2642.2
	Piecewise Weibull	0.0	0.0419	0.784	0.475	**2626.1**
D1	Exponential	–	0.0257	–	0.464	3797.5
	Weibull	–	0.0257	0.768	0.495	3760.6
	Piecewise exponential	12.3	0.0257	–	0.276	3736.6
	Piecewise Weibull	12.3	0.0248	0.895	0.314	**3732.1**
D2	Exponential	–	0.0259	–	0.580	3954.7
	Weibull	–	0.0258	0.806	0.605	3929.2
	Piecewise exponential	8.2	0.0263	–	0.448	3927.4
	Piecewise Weibull	8.2	0.0252	0.887	0.507	**3921.3**
E	Exponential	–	0.0928	–	0.535	1846.7
	Weibull	–	0.0994	0.866	0.572	**1839.1**
	Piecewise exponential	0.0	0.0928	–	0.535	1846.7
	Piecewise Weibull	0.0	0.0994	0.866	0.572	**1839.1**
G	Exponential	–	0.1234	–	0.605	2273.7
	Weibull	–	0.1280	0.908	0.634	2270.2
	Piecewise exponential	3.3	0.1182	–	0.539	**2260.2**
	Piecewise Weibull	3.3	0.1187	1.019	0.522	2262.1
H	Exponential	–	0.0684	–	0.618	4139.7
	Weibull	–	0.0709	0.905	0.645	4132.7
	Piecewise exponential	6.6	0.0683	–	0.483	**4109.3**
	Piecewise Weibull	6.6	0.0683	1.005	0.480	4111.3

- Proportional hazards models were adequate only when no delay was present.

These findings reinforce the Trial G conclusion that explicitly modelling delayed onset improves descriptive accuracy and prevents attenuation of the estimated late treatment benefit.

7.5.3 Question 3: Interim Decision Analysis

Table 7.8: *Interim stopping behaviour for overall survival (OS) across all trials under one-sided O'Brien–Fleming ($\alpha = 0.025$) efficacy spending and Kim–DeMets ($\beta = 0.10$) futility spending. The first boundary crossing is reported for each rule, with futility treated as non-binding in this diagnostic assessment. Percentages indicate the proportion of total observed events at crossing. Interpretations follow the terminology introduced in Section 7.4.3.*

Trial	Futility spending	Futility crossed	Efficacy crossed	Interpretation
A	$\gamma = 0.75$	26 (6.1%)	298 (70.0%)	Futility stop consistent with final outcome
	$\gamma = 2.5$	388 (91.1%)	298 (70.0%)	Premature efficacy stop (no benefit)
	$\gamma = 6$	399 (93.7%)	298 (70.0%)	Premature efficacy stop (no benefit)
B	$\gamma = 0.75$	–	170 (25.3%)	Efficacy stop consistent with final outcome
	$\gamma = 2.5$	–	170 (25.3%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	170 (25.3%)	Efficacy stop consistent with final outcome
C	$\gamma = 0.75$	–	108 (36.6%)	Efficacy stop consistent with final outcome
	$\gamma = 2.5$	–	108 (36.6%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	108 (36.6%)	Efficacy stop consistent with final outcome
D1	$\gamma = 0.75$	–	209 (53.6%)	Efficacy stop consistent with final outcome
	$\gamma = 2.5$	–	209 (53.6%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	209 (53.6%)	Efficacy stop consistent with final outcome
D2	$\gamma = 0.75$	–	250 (60.7%)	Efficacy stop consistent with final outcome
	$\gamma = 2.5$	–	250 (60.7%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	250 (60.7%)	Efficacy stop consistent with final outcome
E	$\gamma = 0.75$	–	145 (56.6%)	Efficacy stop consistent with final outcome
	$\gamma = 2.5$	–	145 (56.6%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	145 (56.6%)	Efficacy stop consistent with final outcome
G	$\gamma = 0.75$	59 (17.4%)	196 (57.8%)	Premature futility stop (benefit later emerged)
	$\gamma = 2.5$	–	196 (57.8%)	Efficacy stop consistent with final outcome
	$\gamma = 6$	–	196 (57.8%)	Efficacy stop consistent with final outcome
H	$\gamma = 0.75$	38 (7.0%)	373 (69.1%)	Premature futility stop (benefit later emerged)
	$\gamma = 2.5$	98 (18.1%)	373 (69.1%)	Premature futility stop (benefit later emerged)
	$\gamma = 6$	–	373 (69.1%)	Efficacy stop consistent with final outcome

The risk of premature futility stopping under delayed effects was evaluated using retrospective group-sequential analyses. Results in Table 7.8 (OS) and Table C.3 (PFS) show that:

- Trials with larger delays (A, G, H) showed materially increased rates of *premature futility stop*, especially under aggressive spending rules.
- Conservative futility boundaries substantially reduced this risk while still enabling appropriate efficacy decisions.
- Trials without significant delay exhibited monitoring behaviour consistent with design expectations.

Methodologically, these results demonstrate that monitoring strategies assuming proportional hazards may incorrectly interpret early data as evidence of futility when the true treatment benefit emerges only after a delay.

7.5.4 Question 4: Power Loss

Delayed effects had a substantial impact on empirical log-rank power relative to design-stage expectations. When the proportional hazards (PH) assumption was retained but the observed delay imposed, several trials (A, D1, D2, G, H) exhibited power reductions exceeding 25–35 percentage points (Table 7.9). These losses reflect the accumulation of early, uninformative events during the pre-benefit period, which attenuate the treatment signal in standard time-to-event analyses.

In contrast, for trials where the observed delay coincided with a stronger-than-anticipated post-delay hazard reduction (e.g., Trials C and E), empirical power was maintained or even increased relative to the original design. Two consistent features characterised these instances:

1. Lower baseline event rates, implying slower accumulation of information and reduced sensitivity to the timing of treatment effect onset; and
2. A treatment effect that exceeded the value assumed in the design.

Together, these factors counteracted the dilution of treatment information caused by the delay, resulting in apparent gains in operating characteristics.

Overall, these results highlight that the impact of delayed effects on power is highly context dependent. Designs assuming proportional hazards may perform adequately when delays are short, baseline hazards are low, or post-delay effects are strong, but can suffer severe efficiency loss otherwise. From a design perspective, these findings reinforce that delay-induced mis-specification can jeopardise confirmatory success unless mitigated through larger sample sizes, delayed primary analyses, or explicit piecewise statistical modelling.

7.6 Discussion

This chapter has demonstrated the application of the delayed-treatment-effect (DTE) framework developed earlier in the thesis to seven completed oncology trials. Access to individual participant data enabled a consistent evaluation of survival behaviour, parametric model performance, interim monitoring implications, and design-stage operating characteristics in real clinical settings.

Across multiple studies, delayed separation of survival curves was both empirically evident and clinically plausible. In trials where treatment benefit emerged only after a delay period, proportional hazards (PH) modelling did not capture the evolving hazard structure. Piecewise parametric models provided superior descriptive accuracy and yielded more realistic estimates of treatment benefit, particularly at later follow-up times. These observations reinforce the methodological conclusions of Chapters 4 and 5: non-proportional hazards should be regarded as a common and expected feature of contemporary immuno-oncology trials rather than an exception.

An important advantage of analysing real clinical trial data is that it incorporates the complexities of patient flow and site activation that stylised simulations cannot fully reflect. Recruitment is influenced by centre opening schedules, referral pathways, and temporal shifts in patient characteristics. As a result, early and late enrolees may differ systematically in ways that affect event timing and hazard evolution. Real-world datasets therefore provide a richer and more realistic basis for evaluating the practical implications of delayed effects than hypothetical scenarios alone.

The behaviour of early estimates is also consistent with findings from Edwards et al., 2023, who examined the stability of treatment effect estimates for continuous and binary outcomes and observed approximate stabilisation once two-thirds of the total information had accrued. In the present context, stabilisation occurs even later due to the presence of a delay: early events largely reflect the pre-benefit period, causing the observed hazard ratio to fluctuate markedly until well into follow-up. This has direct operational consequences—an interim futility assessment conducted during this phase is more likely to underestimate the eventual treatment effect.

The case studies further showed that delayed effects can materially distort interim monitoring decisions. When early data are dominated by the pre benefit period, conventional proportional hazards based futility boundaries may lead to premature termination of a trial that would ultimately demonstrate efficacy. Such patterns were evident in several studies. Approaches that reduce the influence of early, non informative hazards, such as more conservative futility spending, deferral of interim analyses, or the use of predictive probability monitoring as developed in Chapter 5, offer improved alignment between interim decisions and long term outcomes.

The efficiency of event-driven designs is similarly sensitive to the timing of treatment effect emergence. PH-based design calculations treat all events as equally informative, but when a delay is present, a substantial proportion of events occur before the treatment effect manifests. This dilutes the test statistic and reduces the probability of detecting a true effect at the planned analysis. Empirical power losses of 25–35 percentage points were observed in several trials when the delay-time was ignored. These findings highlight a vulnerability of PH-based planning: nominal power and assurance may be substantially overstated unless assumptions about the length of delay are incorporated explicitly.

Several limitations of the analyses should be acknowledged. Owing to anonymisation procedures, absolute calendar times were unavailable, precluding reconstruction of accrual dynamics, dropout profiles, or the original interim analysis timings. The case studies were intentionally selected based on evidence of delayed separation, introducing the potential for publication or selection bias when considering the broader prevalence of delayed effects in oncology. Moreover, the analyses were exploratory rather than confirmatory; clinical interpretation and contextualisation remain the responsibility of the original trial investigators.

Despite these limitations, the case studies provide strong empirical support for the methodological developments presented in this thesis. Designing trials solely under PH assumptions risks inefficiency, premature termination, and incorrect inference when treatment effects emerge gradually. Explicitly recognising and planning for a potential delay period is therefore essential when clinical benefit may depend on biological activation, immune priming, or prolonged disease stabilisation.

The final chapter integrates insights from both the theoretical developments and the empirical case studies to formulate practical recommendations for the design, monitoring, and analysis of future survival trials in which delayed treatment effects are anticipated.

Table 7.9: *Estimated power across all trials under three design scenarios. For trials where a nonzero delay was identified, three rows are shown: (i) the original proportional hazards (PH) design assumption; (ii) the best-fitting delayed-effect model estimated from data; and (iii) the original PH design assumptions with the observed delay applied. For trials with no estimated delay, only two rows are reported since scenarios (i) and (iii) are equivalent.*

		Design Parameters					Operating Characteristics	
Trial	**Scenario**	**Delay**	$\boldsymbol{\lambda_c}$	$\boldsymbol{\gamma_c}$	HR*	**Recruit.**	**Power (%)**	**Duration**
A	*(i)*	0.0	0.0495	–	0.75	31.5	69	35
	(ii)	5.3	0.0420	0.923	0.794	31.5	21	36
	(iii)	5.3	0.0495	–	0.75	31.5	20	34
B	*(i)* or *(iii)*	0.0	0.0468	–	0.76	17.4	90	38
	(ii)	0.0	0.0303	–	0.743	17.4	95	54
C	*(i)* or *(iii)*	0.0	0.0693	–	0.69	13.2	90	32
	(ii)	0.0	0.0419	0.784	0.475	13.2	100	70
D1	*(i)*	0.0	0.0856	–	0.72	8.1	94	38
	(ii)	12.3	0.0248	0.895	0.314	8.1	100	104
	(iii)	12.3	0.0856	–	0.72	8.1	29	35
D2	*(i)*	0.0	0.0856	–	0.72	8.1	94	38
	(ii)	8.2	0.0252	0.887	0.507	8.1	100	85
	(iii)	8.2	0.0856	–	0.72	8.1	50	36
E	*(i)* or *(iii)*	0.0	0.0488	–	0.61	14.2	92	24
	(ii)	0.0	0.0994	0.866	0.572	14.2	96	26
G	*(i)*	0.0	0.1155	–	0.67	14.2	90	25
	(ii)	3.3	0.1182	–	0.54	14.2	84	26
	(iii)	3.3	0.1155	–	0.67	14.2	49	24
H	*(i)*	0.0	0.0866	–	0.72	13.1	90	24
	(ii)	6.6	0.0683	–	0.483	13.1	93	29
	(iii)	6.6	0.0866	–	0.72	13.1	56	23

Chapter 8

Conclusion

8.1 Introduction and Motivation

This thesis has addressed a methodological gap in the design of time-to-event clinical trials in which therapeutic benefit is not immediate but emerges only after a delay period. Such *delayed treatment effects* (DTEs) are increasingly observed in immuno-oncology and other areas where complex biological mechanisms give rise to non-proportional hazards (NPH) (Mukhopadhyay et al., 2022). Although these patterns are well documented, prevailing approaches to trial design and interim monitoring continue to rely on proportional hazards assumptions and fixed, instantaneous effect sizes. As a result, standard design tools can yield misleading operating characteristics—both overstating power at the planning stage and encouraging premature futility stopping during trial execution.

Traditional power calculations require investigators to specify a single assumed treatment effect, an approach that is particularly fragile when uncertainty surrounds both the magnitude and timing of benefit (Jones et al., 2003). Bayesian assurance offers a principled alternative by integrating over a prior distribution for unknown parameters and yielding a probability of trial success that explicitly acknowledges design uncertainty (O'Hagan, Stevens, et al., 2005). When empirical evidence is sparse, structured expert elicitation provides a rigorous means of quantifying uncertainty in clinically meaningful terms (Garthwaite et al., 2005). The increasing prominence of adaptive designs further motivates investigation into how assurance, augmented with elicited priors, can guide interim analyses in the presence of delayed effects, thereby reducing the risk of discarding effective therapies on the basis of uninformative early data (Pallmann et al., 2018).

Together, these considerations motivated the development of a coherent methodological

framework for incorporating delayed effects into trial design, interim monitoring, and decision-making—culminating in the central research question of this thesis.

8.2 Research Question Revisited

This thesis was guided by the overarching question:

> *"How can adaptive time-to-event trials be designed to incorporate elicited prior information about delayed treatment effects and to quantify probability of success under these assumptions?"*

The results demonstrate that addressing this question requires more than incremental modification of existing design components. Instead, it necessitates an integrated framework that links expert elicitation, coherent prior construction, assurance-based design evaluation, and adaptive interim monitoring. The thesis develops such a framework and demonstrates its practical relevance through empirical analyses and a fully implemented software platform.

8.3 Addressing the Research Objectives

The five research objectives set out in Section 1.1 are revisited here, summarising how each has been addressed in the thesis.

8.3.1 Objective 1: Develop an elicitation framework tailored to delayed treatment effects

The thesis introduced a structured elicitation framework for capturing expert beliefs about key parameters governing delayed treatment effect trials, including the control-group hazard, the delay duration, and the post-delay hazard ratio. The elicitation protocol specifies the questions posed to experts and formalises how these responses map to the parameters of interest in the underlying survival model.

8.3.2 Objective 2: Extend assurance methodology to incorporate elicited priors

The assurance framework was extended beyond point-effect assumptions to settings in which treatment benefit emerges over time. By integrating elicited priors into analytic and simulation-based evaluations, the thesis developed probability-of-success measures that reflect uncertainty in both the magnitude and timing of treatment effects. This enables feasibility assessments and design justifications that are more realistic than those based on fixed, optimistic assumptions.

8.3.3 Objective 3: Adapt assurance for adaptive time-to-event designs

The methods were extended to group-sequential and other adaptive designs, yielding delay-aware predictive probabilities and interim decision rules. These developments address the susceptibility of interim analyses to misleading early data when treatment effects are delayed. Within this framework, assurance is used to evaluate and optimise adaptive design choices based on their resulting operating characteristics.

8.3.4 Objective 4: Evaluate the practical implications using real clinical trial data

Analyses of anonymised oncology trials provided empirical evidence of delayed treatment effects across multiple indications. Piecewise hazard models frequently outperformed proportional hazards specifications, with estimated delay times consistent with plausible biological mechanisms. Interim monitoring evaluations showed that ignoring delay can lead to substantial power loss and premature futility stopping, reinforcing the need for design methods that explicitly account for delayed effects.

8.3.5 Objective 5: Implement the methodology in an open-source software package

The `DTEAssurance` R package and accompanying Shiny applications operationalise the proposed methodology, supporting elicitation, prior construction, assurance calculation, adaptive design simulation, and visualisation of operating characteristics. This implementation promotes reproducibility, facilitates adoption, and enables applied trialists to incorporate delayed-effect considerations into routine design practice.

8.4 Impact and Broader Context

8.4.1 Methodological and Scientific Impact

This thesis contributes to a methodological shift in clinical trial design: from fixed, deterministic assumptions to explicit representation of uncertainty. By extending assurance to survival trials with delayed treatment effects, the work demonstrates how Bayesian reasoning can be embedded in design justification and decision-making. Rather than evaluating power at a single assumed effect size, the assurance-based approach evaluates the likelihood of trial success given current knowledge—an inherently more realistic and defensible perspective.

The elicitation component advances the role of expert judgement in trial design. In areas where empirical data are sparse, the ability to translate qualitative clinical insight into quantitative prior distributions promotes a more transparent and accountable design process. This aligns with evolving regulatory expectations emphasising traceability of assumptions and reproducibility of design rationale (U.S. Food and Drug Administration, 2019).

The work also contributes to adaptive design methodology. Interim monitoring is highly sensitive to delayed effects, and early data can underestimate true benefit. By incorporating delay uncertainty into predictive probabilities and stopping rules, the thesis supports more reliable interim decision-making and reduces the risk of prematurely discarding effective therapies. This adds to ongoing conversations in the adaptive design literature on balancing flexibility with statistical rigour (Pallmann et al., 2018; Berry, 2006).

8.4.2 Practical Impact and Software Implementation

The `DTEAssurance` R package provides an accessible platform for applying the methods developed in this thesis. Through interactive visualisation and streamlined workflows, the software brings elicitation, assurance computation, and adaptive design within reach of practitioners who may not have extensive Bayesian training. By lowering these technical barriers, the methodology becomes feasible for use in early engagement discussions, design justification documents, and regulatory submissions.

8.4.3 Implications for Clinical Research

Neglecting delayed effects can have substantial consequences for trial efficiency, interpretation, and ethical conduct. The empirical analyses demonstrated that delays of 2–8 months were common across the case studies, with clear consequences: power losses of 25–65 percentage points when delays were ignored in design assumptions, and premature futility stopping at 7–18% information under aggressive monitoring schemes. These findings highlight the importance of designing trials that reflect the complexity of treatment response dynamics and ensure that patients, resources, and time are used responsibly.

8.5 Limitations and Future Directions

While this thesis offers methodological, practical, and software contributions, several limitations remain and point toward promising directions for future research.

8.5.1 Expert Elicitation

The assurance framework relies on elicited priors, and although structured protocols such as SHELF provide rigour, expert judgements inevitably carry subjectivity and potential bias. Eliciting joint beliefs about delay, effect size, and event rates can be cognitively demanding, and this work did not evaluate elicited priors against realised trial outcomes or address how to reconcile divergent expert opinions.

Future research should examine the empirical calibration of expert judgements by comparing elicited priors with observed outcomes across multiple trials. As external evidence accumulates, elicited priors can be updated formally using Bayes' theorem, yielding posteriors that combine expert judgement with observed data.

8.5.2 Model Parameterisation and General Non-Proportional Hazards

The modelling framework adopts a single change point with a transition to a post-delay hazard ratio. Although this suffices to illustrate core ideas, real treatment effects may emerge gradually, diminish over time, or follow more complex hazard dynamics. Extending the framework to include smooth transitions or multiple change points—drawing on flexible survival models such as Royston–Parmar or spline-based

hazards—may improve realism but raises challenges for elicitation and assurance computation.

More broadly, delayed benefit is only one form of non-proportional hazards. Patterns such as waning effects, early harm, or crossing hazards also warrant more complex design strategies. A future programme of research could develop a taxonomy of NPH scenarios with corresponding elicitation templates and design tools, examining the trade-offs between model flexibility, elicitation feasibility, and computational tractability.

8.5.3 Empirical Validation

The empirical evaluation focused on a selected sample of trials exhibiting visually apparent delayed separation, which introduces selection bias and precludes inference about the broader prevalence of DTEs. Analyses were retrospective and constrained by de-identified data, preventing reconstruction of true accrual profiles or original interim monitoring schedules.

Unbiased estimation of DTE prevalence would require a systematic review of completed trials selected without reference to hazard-pattern appearance. Prospective validation—applying these design tools in ongoing trials—will be essential for assessing operational feasibility, regulatory acceptability, and the practical utility of elicitation-informed assurance. Engagement with trial sponsors and regulators will be crucial to understanding expectations around elicitation documentation and assurance thresholds.

8.5.4 Software Development and Translation to Practice

The `DTEAssurance` package supports the key methodological components of this thesis, but continued development is needed to incorporate more flexible hazard models, improve computational efficiency, and expand visualisation tools for communicating assurance and interim-monitoring implications. Case studies documenting prospective, real-world use of the package would help bridge the gap between methodological development and routine practice.

8.6 Concluding Remarks

This thesis has addressed the challenge of designing adaptive time-to-event trials in settings where treatment benefits emerge only after a delay. By integrating Bayesian assurance, structured expert elicitation, and adaptive monitoring within a framework, it contributes to an evolving shift in trial design toward explicit modelling of uncertainty and realistic representation of treatment dynamics.

The methodological developments establish how assurance can provide a coherent basis for evaluating trial success when key parameters—particularly the timing and magnitude of treatment benefit—are uncertain. The elicitation framework allows expert beliefs about delayed effects to be translated into joint prior distributions, offering a principled alternative to ad hoc assumptions when empirical evidence is limited. Extending these ideas to adaptive designs demonstrates how delay-informed predictive probabilities can reduce the risk of premature futility stopping—an issue of practical importance in trials where early data may not yet reflect true treatment benefit.

Empirical analyses of seven completed oncology trials reinforce the practical relevance of these methodological concerns. Delayed effects were common, and ignoring them in design assumptions led to substantial power loss and increased risk of inappropriate early termination. These findings underscore that delayed treatment effects are not isolated anomalies but recurring features of modern therapeutic settings that require dedicated design strategies.

The accompanying `DTEAssurance` package ensures that the methodological contributions are accessible to practitioners. By providing tools for elicitation, assurance computation, and visualisation, the software supports uncertainty-aware design without requiring deep familiarity with Bayesian computation. This practical implementation bridges the gap between methodological development and real-world application.

Looking ahead, this work points toward a design paradigm that treats uncertainty as an inherent feature of clinical trials to be characterised rather than minimised. For survival endpoints with delayed or time-varying effects, this entails adopting models that reflect biological complexity, designing studies that are robust to plausible departures from idealised assumptions, and ensuring that design assumptions are transparent and justifiable. Further progress will require extensions to more flexible hazard structures, prospective validation in ongoing trials, and continued engagement with regulators to support the use of elicitation- and assurance-based design approaches.

The methodological framework, empirical analyses, and software tools presented here provide a foundation for such developments. If these contributions support more reliable detection of treatment benefit in settings where effects emerge gradually, they will have met their intended aims.

Bibliography

Abbas, A. E., D. V. Budescu, H.-T. Yu, and R. Haggerty (2008). "A Comparison of Two Probability Encoding Methods: Fixed Probability vs. Fixed Variable Values". In: *Decision Analysis* 5.4, pp. 190–202. DOI: 10.1287/deca.1080.0126. eprint: https://doi.org/10.1287/deca.1080.0126 (cit. on p. 28).

Alhussain, Z. and J. Oakley (2020). "Assurance for clinical trial design with normally distributed outcomes: eliciting uncertainty about variances". In: *Pharmaceutical Statistics* 19(6), pp. 827–839. DOI: 10.1002/pst.2040 (cit. on pp. 13, 20, 108).

Anderson, K. (2025). *gsDesign: Group Sequential Design*. R package version 3.7.0 (cit. on pp. 94, 100).

Armitage, P., C. McPherson, and B. Rowe (1969). "Repeated significance tests on accumulating data". In: *Journal of the Royal Statistical Society: Series A (General)* 132.2, pp. 235–244 (cit. on p. 93).

Arnold, B. F., D. R. Hogan, J. M. J. Colford, and A. E. Hubbard (2011). "Simulation methods to estimate design power: an overview for applied research". In: *BMC Medical Research Methodology* 11.1, p. 94. DOI: 10.1186/1471-2288-11-94 (cit. on p. 15).

Azzolina, D., P. Berchialla, D. Gregori, and I. Baldi (2021). "Prior Elicitation for Use in Clinical Trial Design and Analysis: A Literature Review". In: *International Journal of Environmental Research and Public Health* 18.4, p. 1833. DOI: 10.3390/ijerph18041833 (cit. on p. 44).

Bardo, M., C. Huber, N. Benda, J. Brugger, T. Fellinger, Vaidotas Galaune, J. Heinz, H. Heinzl, A. C. Hooker, Florian Klinglmüller, Franz König, T. Mathes, M. Mittlböck, M. Posch, R. Ristl, and T. Friede (2024). "Methods for non-proportional hazards in clinical trials: A systematic review". In: *Statistical methods in medical research* 33.6. Publisher: SAGE Publishing, pp. 1069–1092. DOI: 10.1177/09622802241242325 (cit. on p. 52).

Bauer, P. and K. Köhne (1994). "Evaluation of experiments with adaptive interim analyses". In: *Biometrics*, pp. 1029–1041 (cit. on pp. 105, 110).

Ben-Eltriki, M., A. Rafiq, A. Paul, D. Prabhu, M. O. Afolabi, R. Baslhaw, C. J. Neilson, M. Driedger, S. M. Mahmud, T. Lacaze-Masmonteil, et al. (2024). "Adaptive designs in clinical trials: a systematic review-part I". In: *BMC Medical Research Methodology* 24.1, p. 229 (cit. on p. 81).

Berry, D. A. (1996). *Statistics: a Bayesian perspective*. Vol. 4. Duxbury Press Belmont, CA (cit. on p. 11).

Berry, D. A. (2006). "Bayesian clinical trials". In: *Nature reviews Drug discovery* 5.1, pp. 27–36 (cit. on pp. 81, 170).

Berry, D. A. (1993). "A Case for Bayesianism in Clinical Trials". In: *Statistical Science* 8.1, pp. 39–44. DOI: 10.1214/ss/1177010337 (cit. on p. 9).

Bertsche, A., F. Fleischer, J. Beyersmann, and G. Nehmiz (2019). "Bayesian Phase II optimization for time-to-event data based on historical information". In: *Statistical Methods in Medical Research* 28.4, pp. 1272–1289 (cit. on pp. 55, 65, 198).

Best, N., N. Dallow, and T. Montague (2020). "Prior Elicitation". In: *Bayesian Methods in Pharmaceutical Research*. Ed. by D. A. Berry et al. Chapman and Hall/CRC, pp. 87–109 (cit. on pp. 42, 44).

Bhatt, D. L. and C. Mehta (2016). "Adaptive designs for clinical trials". In: *New England Journal of Medicine* 375.1, pp. 65–74 (cit. on p. 82).

Bolliger, D., M. Seeberger, G. Buse, P. Christen, L. Gürke, and M. Filipovic (2007). "Randomized clinical trial of moxonidine in patients undergoing major vascular surgery". In: *The British journal of surgery* 94, pp. 1477–84. DOI: 10.1002/bjs.6012 (cit. on p. 17).

Borghaei, H., S. Gettinger, et al. (2021). "Five-Year Outcomes From the Randomized, Phase III Trials CheckMate 017 and 057: Nivolumab Versus Docetaxel in Previously Treated Non–Small-Cell Lung Cancer". In: *Journal of Clinical Oncology* 39.7, pp. 723–733. DOI: 10.1200/jco.20.01605 (cit. on pp. xv, 73, 209).

Borghaei, H., L. Paz-Ares, L. Horn, D. R. Spigel, M. Steins, N. E. Ready, L. Q. Chow, E. E. Vokes, E. Felip, E. Holgado, et al. (2015). "Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer". In: *New England Journal of Medicine* 373.17, pp. 1627–1639 (cit. on p. 209).

Bothwell, L. E., J. Avorn, N. F. Khan, and A. S. Kesselheim (2018). "Adaptive design clinical trials: a review of the literature and ClinicalTrials. gov". In: *BMJ open* 8.2, e018320 (cit. on pp. 102, 110).

Brahmer, J., K. L. Reckamp, P. Baas, et al. (2015). "Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer". In: *New England Journal of Medicine* 373.2, pp. 123–135. DOI: 10.1056/nejmoa1504627 (cit. on pp. xv, xix, 49, 73, 74, 209).

Brahmer, J. R., J.-S. Lee, T.-E. Ciuleanu, R. Bernabe Caro, M. Nishio, L. Urban, C. Audigier-Valette, L. Lupinacci, R. Sangha, A. Pluzanski, et al. (2023). "Five-year survival outcomes with nivolumab plus ipilimumab versus chemotherapy as first-line treatment for metastatic non–small-cell lung cancer in CheckMate 227". In: *Journal of Clinical Oncology* 41.6, pp. 1200–1212 (cit. on p. 209).

Bretz, F., F. Koenig, W. Brannath, E. Glimm, and M. Posch (2009). "Adaptive designs for confirmatory clinical trials". In: *Statistics in medicine* 28.8, pp. 1181–1217 (cit. on p. 82).

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng (2011). *Handbook of markov chain monte carlo*. CRC press (cit. on p. 89).

Burton, A., D. G. Altman, P. Royston, and R. L. Holder (2006). "The design of simulation studies in medical statistics". In: *Statistics in Medicine* 25.24, pp. 4279–4292. DOI: https://doi.org/10.1002/sim.2673. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2673 (cit. on p. 15).

Cetinyurek Yavuz, A., M. B. N. Fayyad, C. Jiang, F. Brion Bouvier, C. Beji, S. Zebachi, G. Y. Hayek, B. Amzal, R. Porcher, J. Tanniou, et al. (2025). "On the Concepts, Methods, and Use of "Probability of Success" for Drug Development Decision-Making: A Scoping Review". In: *Clinical Pharmacology & Therapeutics* 117.4, pp. 967–977 (cit. on p. 44).

Chaloner, K. and F. S. Rhame (2001). "Quantifying and documenting prior beliefs in clinical trials". In: *Statistics in Medicine* 20.4, pp. 581–600. DOI: https://doi.org/10.1002/sim.694. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.694 (cit. on p. 22).

Chang, W., J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges (2025). *shiny: Web Application Framework for R*. R package version 1.11.1. DOI: 10.32614/CRAN.package.shiny (cit. on p. 5).

Chen, T.-T. (2013). "Statistical issues and challenges in immuno-oncology". In: *Journal for ImmunoTherapy of Cancer* 1.1. DOI: 10.1186/2051-1426-1-18 (cit. on pp. 2, 50).

Chow, S.-C., M. Chang, and A. Pong (2005). "Statistical consideration of adaptive methods in clinical development". In: *Journal of biopharmaceutical statistics* 15.4, pp. 575–591 (cit. on p. 110).

Chow, S.-C. and R. Corey (2011). "Benefits, challenges and obstacles of adaptive clinical trial designs". In: *Orphanet journal of rare diseases* 6, pp. 1–10 (cit. on p. 104).

Chuang-Stein, C. (2006). "Sample size and the probability of a successful trial". In: *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 5.4, pp. 305–309 (cit. on p. 10).

Clemen, R. T. and R. L. Winkler (1999). "Combining probability distributions from experts in risk analysis". In: *Risk analysis* 19, pp. 187–203 (cit. on p. 32).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. routledge (cit. on p. 11).

Collett, D. (2015). *Modelling Survival Data in Medical Research*. 3rd. Boca Raton, FL: CRC Press (cit. on p. 52).

Collignon, O., A. Schiel, C.-F. Burman, K. Rufibach, M. Posch, and F. Bretz (2022). "Estimands and complex innovative designs". In: *Clinical Pharmacology & Therapeutics* 112.6, pp. 1183–1190 (cit. on p. 102).

Colson, A. R. and R. M. Cooke (2018). "Expert elicitation: using the classical model to validate experts' judgments". In: *Review of Environmental Economics and Policy* (cit. on p. 38).

Cooke, R. (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford university press (cit. on pp. 33, 38).

Cooke, R. M. and D. Solomatine (1992). "EXCALIBR Integrated System for Processing Expert Judgements version 3.0". In: *Delft University of Technology and SoLogic Delft, Delft* (cit. on p. 38).

Crisp, A., S. Miller, D. Thompson, and N. Best (2018). "Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development". In: *Pharmaceutical Statistics* 17.4, pp. 317–328 (cit. on p. 21).

Cytel (2025). *East Version 6.4*. Software for clinical trial design and analysis. Cytel Inc. (cit. on p. 100).

Dalkey, N. and O. Helmer (1963). "An experimental application of the Delphi method to the use of experts". In: *Management science* 9.3, pp. 458–467 (cit. on pp. 34, 37).

Dallow, N., N. Best, and T. H. Montague (2018). "Better decision making in drug development through adoption of formal prior elicitation". In: *Pharmaceutical Statistics* 17.4, pp. 301–316. DOI: 10.1002/pst.1854 (cit. on pp. 1, 42, 43).

Daneshkhah, A. and J. E. Oakley (2010). "Eliciting multivariate probability distributions". In: *Rethinking Risk Measurement and Reporting: Volume 1*. Ed. by K. Böcker. London: Risk Books (cit. on p. 31).

DeMets, D. L. (2006). "Futility approaches to interim monitoring by data monitoring committees". In: *Clinical Trials* 3.6, pp. 522–529 (cit. on pp. 85, 91).

Demets, D. L. and K. K. G. Lan (1994). "Interim analysis: The alpha spending function approach". In: *Statistics in Medicine* 13.13-14, pp. 1341–1352. DOI: 10.1002/sim.4780131308 (cit. on pp. 104, 105).

Dent, L. and J. Raftery (2011). "Treatment success in pragmatic randomised controlled trials: a review of trials funded by the UK Health Technology Assessment programme". In: *Trials* 12, pp. 1–10 (cit. on p. 101).

Dias, L. C., A. Morton, and J. Quigley (2018). "Elicitation: State of the art and science". In: *Elicitation: The science and art of structuring judgement*, pp. 1–14 (cit. on p. 2).

Dimairo, M., S. A. Julious, S. Todd, J. P. Nicholl, and J. Boote (2015). "Cross-sector surveys assessing perceptions of key stakeholders towards barriers, concerns and facilitators to the appropriate use of adaptive designs in confirmatory trials". In: *Trials* 16.1, p. 585 (cit. on p. 110).

Dmitrienko, A. and M.-D. Wang (2006). "Bayesian predictive approach to interim monitoring in clinical trials". In: *Statistics in Medicine* 25.13, pp. 2178–2195. DOI: 10.1002/sim.2204 (cit. on p. 85).

Dolan, J. G., D. R. Bordley, and A. I. Mushlin (1986). "An Eualuation of Clinicians' Subjective Prior Probability Estimates". In: *Medical Decision Making* 6.4. PMID: 3773651, pp. 216–223. DOI: 10.1177/0272989X8600600406. eprint: https://doi.org/10.1177/0272989X8600600406 (cit. on p. 22).

Edwards, J. M., S. J. Walters, and S. A. Julious (2023). "A retrospective analysis of conditional power assumptions in clinical trials with continuous or binary endpoints". In: *Trials* 24.1, p. 215 (cit. on pp. 104, 163).

Emerson, S. S., J. M. Kittelson, and D. L. Gillen (2007). "Frequentist evaluation of group sequential clinical trial designs". In: *Statistics in medicine* 26.28, pp. 5047–5080 (cit. on p. 93).

European Food Safety Authority (2014). "Guidance on expert knowledge elicitation in food and feed safety risk assessment". In: *EFSA Journal* 12.6, p. 3734. DOI: 10.2903/j.efsa.2014.3734 (cit. on pp. 36, 39).

European Medicines Agency (2007). *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design*. CHMP/EWP/2459/02, Accessed July 14, 2025 (cit. on pp. 102, 105, 110).

Falconer, J. R., E. Frank, D. L. Polaschek, and C. Joshi (2022). "Methods for eliciting informative prior distributions: A critical review". In: *Decision Analysis* 19.3, pp. 189–204 (cit. on p. 27).

Ferris, R. L., G. Blumenschein Jr, J. Fayette, J. Guigay, A. D. Colevas, L. Licitra, K. Harrington, S. Kasper, E. E. Vokes, C. Even, et al. (2016). "Nivolumab for recurrent squamous-cell carcinoma of the head and neck". In: *New England Journal of Medicine* 375.19, pp. 1856–1867 (cit. on pp. 156, 209).

Fine, G. D. (2007). "Consequences of Delayed Treatment Effects on Analysis of Time-to-Event Endpoints". In: *Drug Information Journal* 41.4, pp. 535–539. DOI: 10.1177/009286150704100412 (cit. on pp. 2, 50, 51).

Fleming, T. R. and D. P. Harrington (1981). "A class of hypothesis tests for one and two sample censored survival data". In: *Communications in Statistics* 10.8, pp. 763–794. DOI: 10.1080/03610928108828073 (cit. on pp. 2, 52).

Fleming, T. R., K. Sharples, J. McCall, A. Moore, A. Rodgers, and R. Stewart (2008). "Maintaining confidentiality of interim data to enhance trial integrity and credibility". In: *Clinical Trials* 5.2, pp. 157–167 (cit. on p. 98).

Friedman, L. M. (2015). *Fundamentals of clinical trials*. Springer (cit. on p. 1).

Gallo, P., C. Chuang-Stein, V. Dragalin, B. Gaydos, M. Krams, and J. Pinheiro (2006). "Adaptive designs in clinical drug development—an executive summary of the PhRMA working group". In: *Journal of biopharmaceutical statistics* 16.3, pp. 275–283 (cit. on pp. 104, 105, 110).

Gallo, P., L. Mao, and V. H. Shih (2014). "Alternative views on setting clinical trial futility criteria". In: *Journal of biopharmaceutical statistics* 24.5, pp. 976–993 (cit. on p. 93).

Garon, E. B. et al. (2014). "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". In: *The Lancet* 384.9944, pp. 665–673. DOI: 10.1016/s0140-6736(14)60845-x (cit. on pp. xiv, 65, 66).

Garthwaite, P. H., J. B. Kadane, and A. O'Hagan (2005). "Statistical methods for eliciting probability distributions". In: *Journal of the American statistical Association* 100.470, pp. 680–701 (cit. on pp. 31, 167).

Gasparini, M., L. Di Scala, F. Bretz, and A. Racin-Poon (2013). "Some uses of predictive probability of success in clinical drug development". In: *Epidemiology, Biostatistics, and Public Health* 10.1 (cit. on p. 10).

Gaydos, B., K. M. Anderson, D. Berry, N. Burnham, C. Chuang-Stein, J. Dudinak, P. Fardipour, P. Gallo, S. Givens, R. Lewis, et al. (2009). "Good practices for adaptive clinical trials in pharmaceutical product development". In: *Drug Information Journal* 43.5, pp. 539–556 (cit. on p. 104).

Ghosh, P., R. Ristl, F. König, M. Posch, C. Jennison, H. Götte, A. Schüler, and C. Mehta (2021). "Robust group sequential designs for trials with survival endpoints and delayed response". In: *Biometrical Journal* 64.2, pp. 343–360. DOI: 10.1002/bimj.202000169 (cit. on pp. 111, 112, 130).

Gillett, R. (1994). "An Average Power Criterion for Sample Size Estimation". In: *The Statistician* 43.3, p. 389. DOI: 10.2307/2348574 (cit. on p. 10).

Gore, S. (1987). "Biostatistics and the Medical Research Council". In: *Medical Research Council News* 35, pp. 19–20 (cit. on p. 28).

Gosling, J. P., J. E. Oakley, and A. O'Hagan (2007). "Nonparametric elicitation for heavy-tailed prior distributions". In: *Bayesian Analysis* 2, pp. 693–718 (cit. on p. 29).

Götte, H., J. Xiong, M. Kirchner, H. Demirtas, and M. Kieser (2020). "Optimal decision-making in oncology development programs based on probability of success for phase III utilizing phase II/III data on response and overall survival". In: *Pharmaceutical Statistics* 19.6, pp. 861–881. DOI: 10.1002/pst.2042 (cit. on p. 10).

Grieve, A. P., S. C. Choi, and P. A. Pepple (1991). "Predictive Probability in Clinical Trials". In: *Biometrics* 47.1, pp. 323–330. ISSN: 0006341X, 15410420 (cit. on p. 9).

Grieve, A. P. (2022). *Hybrid Frequentist/Bayesian Power and Bayesian Power in Planning Clinical Trials*. CRC Press LLC. ISBN: 978-1-00-059023-4 (cit. on p. 10).

Grieve, A. P. (2024). "Probability of success and group sequential designs". In: *Pharmaceutical Statistics* 23.2, pp. 185–203. DOI: https://doi.org/10.1002/pst.2346. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pst.2346 (cit. on pp. 20, 24).

Hampson, L. V., B. Bornkamp, B. Holzhauer, J. Kahn, M. R. Lange, W.-L. Luo, G. D. Cioppa, K. Stott, and S. Ballerstedt (2022). "Improving the assessment of the probability of success in late stage drug development". In: *Pharmaceutical Statistics* 21.2, pp. 439–459 (cit. on pp. 10, 21, 24).

Hampson, L. V., B. Holzhauer, B. Bornkamp, J. Kahn, M. R. Lange, W.-L. Luo, P. Singh, S. Ballerstedt, and G. D. Cioppa (2022). "A new comprehensive approach to assess the probability of success of development programs before pivotal trials". In: *Clinical Pharmacology & Therapeutics* 111.5, pp. 1050–1060 (cit. on pp. 10, 21).

Hampson, L. V. and C. Jennison (2013). "Group sequential tests for delayed responses (with discussion)". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75.1, pp. 3–54 (cit. on pp. 103, 104).

Hampson, L. V., J. Whitehead, et al. (2015). "Elicitation of Expert Prior Opinion: Application to the MYPAN Trial in Childhood Polyarteritis Nodosa". In: *PLOS ONE* 10.3. Publisher: Public Library of Science, e0120981–e0120981. DOI: 10.1371/journal.pone.0120981 (cit. on p. 23).

Harari, O., G. Hsu, L. Dron, J. J. Park, K. Thorlund, and E. J. Mills (2021). "Utilizing Bayesian predictive power in clinical trial design". In: *Pharmaceutical Statistics* 20.2, pp. 256–271 (cit. on p. 10).

He, W., X. Cao, and L. Xu (2012). "A framework for joint modeling and joint assessment of efficacy and safety endpoints for probability of success evaluation and optimal dose selection". In: *Statistics in Medicine* 31.5, pp. 401–419 (cit. on p. 10).

Hellmann, M. D., L. Paz-Ares, R. Bernabe Caro, B. Zurawski, S.-W. Kim, E. Carcereny Costa, K. Park, A. Alexandru, L. Lupinacci, E. de la Mora Jimenez, et al. (2019). "Nivolumab plus ipilimumab in advanced non–small-cell lung cancer". In: *New England Journal of Medicine* 381.21, pp. 2020–2031 (cit. on p. 209).

Hemming, V., M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle (2018). "A practical guide to structured expert elicitation using the IDEA protocol". In: *Methods in Ecology and Evolution* 9.1, pp. 169–180 (cit. on pp. 35, 38).

Herbst, R. S., Y. Sun, W. E. Eberhardt, P. Germonpré, N. Saijo, C. Zhou, J. Wang, L. Li, F. Kabbinavar, Y. Ichinose, S. Qin, L. Zhang, B. Biesma, J. V. Heymach, P. Langmuir, S. J. Kennedy, H. Tada, and B. E. Johnson (2010). "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". In: *The Lancet Oncology* 11.7, pp. 619–626. DOI: 10.1016/s1470-2045(10)70132-7 (cit. on pp. xiv, 65, 66).

Herbst, R. S., G. Giaccone, et al. (2020). "Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC". In: *New England Journal of Medicine* 383.14, pp. 1328–1339. DOI: 10.1056/nejmoa1917346 (cit. on pp. 49, 209).

Hiance, A., S. Chevret, and V. Lévy (2009). "A practical approach for eliciting expert prior beliefs about cancer survival in phase III randomized trial". In: *Journal of Clinical Epidemiology* 62.4, 431–437.e2. DOI: 10.1016/j.jclinepi.2008.04.009 (cit. on pp. 42, 44).

Hobbs, B. P., B. P. Carlin, S. J. Mandrekar, and D. J. Sargent (2011). "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials". In: *Biometrics* 67.3, pp. 1047–1056 (cit. on p. 22).

Holzhauer, B., L. V. Hampson, J. P. Gosling, B. Bornkamp, J. Kahn, M. R. Lange, W.-L. Luo, C. Brindicci, D. Lawrence, S. Ballerstedt, and A. O'Hagan (2022). "Eliciting judgements about dependent quantities of interest: The SHeffield ELicitation Framework extension and copula methods illustrated using an asthma case study". In: *Pharmaceutical Statistics* 21.5, pp. 1005–1021. DOI: 10.1002/pst.2212 (cit. on pp. 23, 31, 42, 44).

Hong, S. and L. Shi (2012). "Predictive power to assist phase 3 go/no go decision based on phase 2 data on a different endpoint". In: *Statistics in medicine* 31, pp. 831–43. DOI: 10.1002/sim.4476 (cit. on p. 21).

Ibrahim, J. G. and M.-H. Chen (2000). "Power prior distributions for regression models". In: *Statistical Science*, pp. 46–60 (cit. on p. 22).

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (2019). *Addendum to ICH E9: Statistical Principles for Clinical Trials (R1)*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. Step 4 Guideline (cit. on pp. 52, 102).

Jackson, C. (2016). "flexsurv: A Platform for Parametric Survival Modeling in R". In: *Journal of Statistical Software* 70.8, pp. 1–33. DOI: 10.18637/jss.v070.i08 (cit. on p. 152).

Jacobs, R. A. (1995). "Methods for combining experts' probability assessments". In: *Neural computation* 7.5, pp. 867–888 (cit. on p. 34).

Jassem, J., F. de Marinis, G. Giaccone, A. Vergnenegre, C. H. Barrios, M. Morise, E. Felip, C. Oprean, Y.-C. Kim, Z. Andric, et al. (2021). "Updated overall survival analysis from IMpower110: atezolizumab versus platinum-based chemotherapy in treatment-naive programmed death-ligand 1–selected NSCLC". In: *Journal of Thoracic Oncology* 16.11, pp. 1872–1882 (cit. on p. 209).

Jennison, C. and B. W. Turnbull (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/Crc. ISBN: 978-0-8493-0316-6 (cit. on pp. 2, 84, 91–93, 97, 98, 101, 102, 105, 110, 115).

Jiang, K. (2011). "Optimal sample sizes and go/no-go decisions for phase II/III development programs based on probability of success". In: *Statistics in Biopharmaceutical Research* 3.3, pp. 463–475 (cit. on p. 21).

Jiang, L., F. Yan, P. F. Thall, and X. Huang (2020). "Comparing Bayesian early stopping boundaries for phase II clinical trials". In: *Pharmaceutical statistics* 19.6, pp. 928–939 (cit. on pp. 108, 109).

Jiménez, J. L., V. Stalbovskaya, and B. Jones (2018). "Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects". In: *Pharmaceutical Statistics* 18.3, pp. 287–303. DOI: 10.1002/pst.1923 (cit. on p. 52).

Johnson, S. R., G. A. Tomlinson, G. A. Hawker, J. T. Granton, H. A. Grosbein, and B. M. Feldman (2010). "A valid and reliable belief elicitation method for Bayesian priors". In: *Journal of clinical epidemiology* 63.4, pp. 370–383 (cit. on p. 28).

Jones, S., S. Carley, and M. Harrison (2003). "An introduction to power and sample size estimation". In: *Emergency Medicine Journal* 20.5, pp. 453–458 (cit. on pp. 1, 167).

Julious, S. A. (2009). *Sample Sizes for Clinical Trials*. Chapman and Hall (cit. on p. 11).

Kadane, J. B. and L. J. Wolfson (1998). "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]". In: *The Statistician*

47.1. Publisher: Wiley-Blackwell, pp. 3–19. DOI: 10.1111/1467-9884.00113 (cit. on p. 59).

Kahneman, D. (2011). "Thinking, fast and slow". In: *Farrar, Straus and Giroux* (cit. on p. 40).

Kairalla, J. A., C. S. Coffey, M. A. Thomann, and K. E. Muller (2012). "Adaptive trial designs: a review of barriers and opportunities". In: *Trials* 13, pp. 1–9 (cit. on p. 103).

Kalbfleisch, J. D. and R. L. Prentice (2002). *The statistical analysis of failure time data*. John Wiley & Sons (cit. on p. 1).

Kim, E. S., V. Hirsh, T. Mok, M. A. Socinski, R. Gervais, Y.-L. Wu, L.-Y. Li, C. L. Watkins, M. V. Sellers, E. S. Lowe, Y. Sun, M.-L. Liao, K. Østerlind, M. Reck, A. A. Armour, F. A. Shepherd, S. M. Lippman, and J.-Y. Douillard (2008). "Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial". In: *The Lancet* 372.9652, pp. 1809–1818. DOI: 10.1016/s0140-6736(08)61758-4 (cit. on pp. xiv, 65, 66).

Kim, K. and D. L. Demets (1987). "Design and analysis of group sequential tests based on the type I error spending rate function". In: *Biometrika* 74.1, pp. 149–154 (cit. on p. 96).

Kinnersley, N. and S. Day (2013). "Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study". In: *Pharmaceutical statistics* 12.2, pp. 104–113 (cit. on pp. 42, 44).

Korn, E. L. and B. Freidlin (2018). "Interim Futility Monitoring Assessing Immune Therapies With a Potentially Delayed Treatment Effect". In: *Journal of clinical oncology* 36.23. Publisher: Lippincott Williams & Wilkins, pp. 2444–2449. DOI: 10.1200/jco.2018.77.7144 (cit. on pp. 112–117, 130).

Kundu, M. G., S. Samanta, and S. Mondal (2023). "Review of calculation of conditional power, predictive power and probability of success in clinical trials with continuous, binary and time-to-event endpoints". In: *Health Services and Outcomes Research Methodology*. DOI: 10.1007/s10742-023-00302-5 (cit. on p. 85).

Lachin, J. M. (2009). "Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit". In: *Clinical Trials* 6.6, pp. 565–573 (cit. on p. 91).

Lan, K. K. G. and D. L. DeMets (1983). "Discrete Sequential Boundaries for Clinical Trials". In: *Biometrika* 70.3, pp. 659–663. ISSN: 00063444 (cit. on pp. 92, 98).

Lan, K. G., P. Hu, and M. A. Proschan (2009). "A Conditional Power Approach to the Evaluation of Predictive Power". In: *Statistics in Biopharmaceutical Research* 1.2, pp. 131–136. DOI: 10.1198/sbr.2009.0035 (cit. on p. 85).

Lee, J. J. and D. D. Liu (2008). "A predictive probability design for phase II cancer clinical trials". In: *Clinical trials* 5.2, pp. 93–106 (cit. on pp. 108, 109).

Lehmacher, W. and G. Wassmer (1999). "Adaptive sample size calculations in group sequential trials". In: *Biometrics* 55.4, pp. 1286–1290 (cit. on pp. 102, 110).

Li, B., L. Su, J. Gao, L. Jiang, and F. Yan (2021). "A group sequential design and sample size estimation for an immunotherapy trial with a delayed treatment effect". In: *Statistical Methods in Medical Research* 30.3, pp. 904–915. DOI: 10.1177/0962280220980780 (cit. on pp. 111, 112, 130).

Liu, N., Y. Zhou, and J. J. Lee (2021). "IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves". In: *BMC Medical Research Methodology* 21.1. DOI: 10.1186/s12874-021-01308-8 (cit. on p. 65).

Machin, D., Y. B. Cheung, and M. Parmar (2006). *Survival analysis: a practical approach*. John Wiley & Sons (cit. on p. 1).

Mehta, C. R. and S. J. Pocock (2011). "Adaptive increase in sample size when interim results are promising: a practical guide with examples". In: *Statistics in medicine* 30.28, pp. 3267–3284 (cit. on p. 82).

Morgan, L. M., J. Wason, K. J. Wilson, and N. Wilson (2025). "Comparing Methods of Expert Elicitation for Treatment Effect or Borrowing Parameters in Standard and Rare Disease Clinical Trials: A Systematic Mapping Study". In: *arXiv preprint arXiv:2508.06288* (cit. on p. 42).

Morgan, M. G. (2014). "Use (and abuse) of expert elicitation in support of decision making for public policy". In: *Proceedings of the National academy of Sciences* 111.20, pp. 7176–7184 (cit. on p. 25).

Morita, S., P. F. Thall, and P. Müller (2008). "Determining the effective sample size of a parametric prior". In: *Biometrics* 64.2, pp. 595–602 (cit. on p. 88).

Morris, T. P., I. R. White, and M. J. Crowther (2019). "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* 38.11, pp. 2074–2102. DOI:

https://doi.org/10.1002/sim.8086. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8086 (cit. on p. 15).

Motzer, R. J., B. Escudier, S. George, H. J. Hammers, S. Srinivas, S. S. Tykodi, J. A. Sosman, E. R. Plimack, G. Procopio, D. F. McDermott, et al. (2020). "Nivolumab versus everolimus in patients with advanced renal cell carcinoma: updated results with long-term follow-up of the randomized, open-label, phase 3 CheckMate 025 trial". In: *Cancer* 126.18, pp. 4156–4167 (cit. on p. 209).

Motzer, R. J., B. Escudier, D. F. McDermott, S. George, H. J. Hammers, S. Srinivas, S. S. Tykodi, J. A. Sosman, G. Procopio, E. R. Plimack, et al. (2015). "Nivolumab versus everolimus in advanced renal-cell carcinoma". In: *New England Journal of Medicine* 373.19, pp. 1803–1813 (cit. on p. 209).

Muirhead, R. and A. I. Soaita (2012). "On an Approach to Bayesian Sample Sizing in Clinical Trials". In: DOI: 10.1214/12-IMSCOLL1007 (cit. on p. 13).

Mukhopadhyay, P., J. Ye, K. M. Anderson, S. Roychoudhury, E. H. Rubin, S. Halabi, and R. J. Chappell (2022). "Log-Rank Test vs MaxCombo and Difference in Restricted Mean Survival Time Tests for Comparing Survival Under Nonproportional Hazards in Immuno-oncology Trials: A Systematic Review and Meta-analysis". en. In: *JAMA Oncology* 8.9, p. 1294. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2022.2666 (cit. on pp. 1, 49, 167).

Neuenschwander, B., G. Capkun-Niggli, M. Branson, and D. J. Spiegelhalter (2010). "Summarizing historical information on controls in clinical trials". In: *Clinical Trials* 7.1, pp. 5–18 (cit. on pp. 22, 55, 198).

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons (cit. on p. 205).

O'Brien, P. C. and T. R. Fleming (1979). "A Multiple Testing Procedure for Clinical Trials". In: *Biometrics* 35.3. Publisher: [Wiley, International Biometric Society], pp. 549–556. ISSN: 0006341X, 15410420. DOI: 10.2307/2530245 (cit. on pp. 82, 93, 94, 97).

O'Hagan, A., C. E. Buck, Alireza Daneshkhah, J. Richard Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow (2006). *Uncertain Judgements*. John Wiley & Sons. ISBN: 978-0-470-03330-2 (cit. on pp. 27, 32, 39, 40).

O'Hagan, A., J. W. Stevens, and M. J. Campbell (2005). "Assurance in clinical trial design". In: *Pharmaceutical Statistics* 4(3), pp. 187–201. DOI: 10.1002/pst.175 (cit. on pp. 1, 10, 20, 24, 167).

Oakley, J. (2025). *SHELF: Tools to Support the Sheffield Elicitation Framework*. R package version 1.12.0. DOI: 10.32614/CRAN.package.SHELF (cit. on p. 37).

Oakley, J. E. and A. O'Hagan (2007). "Uncertainty in prior elicitations: a nonparametric approach". In: *Biometrika* 94.2, pp. 427–441 (cit. on p. 29).

Oakley, J. E. and A. O'Hagan (2025). *SHELF: The Sheffield Elicitation Framework (version 4)*. School of Mathematics and Statistics, University of Sheffield (cit. on pp. 31, 37).

Pallmann, P., A. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight Laura and, J. Holmes, A. Mander, L. Odondi, M. Sydes, S. Villar, J. Wason, C. Weir, G. Wheeler, C. Yap, and T. Jaki (2018). "Adaptive designs in clinical trials: Why use them, and how to run and report them". In: *BMC Medicine* 16. DOI: 10.1186/s12916-018-1017-7 (cit. on pp. 2, 81, 167, 170).

Parmar, M. K. B., D. J. Spiegelhalter, L. S. Freedman, and C. S. Committee (1994). "The chart trials: Bayesian design and monitoring in practice". In: *Statistics in Medicine* 13.13-14, pp. 1297–1312. DOI: https://doi.org/10.1002/sim.4780131304. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780131304 (cit. on p. 9).

PHUSE (2020). *Data Anonymisation and Risk Assessment Automation*. p. 10 (cit. on p. 147).

Plummer, M. (2025). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-17. DOI: 10.32614/CRAN.package.rjags (cit. on p. 127).

Pocock, S. J. (1977). "Group sequential methods in the design and analysis of clinical trials". In: *Biometrika* 64.2, pp. 191–199 (cit. on pp. 82, 94, 97).

Pocock, S. J. (1983). *Clinical Trials: A Practical Approach*. Chichester, UK: John Wiley & Sons (cit. on p. 93).

Prior, T. J. (2020). "Group sequential monitoring based on the maximum of weighted log-rank statistics with the Fleming-Harrington class of weights in oncology clinical trials". In: *Statistical Methods in Medical Research* 29.12, pp. 3525–3532. DOI: 10.1177/0962280220931560 (cit. on p. 112).

Proschan, M. A., K. G. Lan, and J. T. Wittes (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer (cit. on pp. 92, 98, 101, 102, 104, 105, 110).

Quinlan, J. A. and M. Krams (2006). "Implementing adaptive designs: logistical and operational considerations". In: *Drug Information Journal* 40.4, pp. 437–444 (cit. on pp. 102, 104, 110).

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on p. 5).

Ren, S. and J. E. Oakley (2014). "Assurance calculations for planning clinical trials with time-to-event outcomes". In: *Statistics in Medicine* 33.1, pp. 31–45 (cit. on pp. 20, 56).

Ribas, A., R. Kefford, M. A. Marshall, C. J. Punt, J. B. Haanen, M. Marmol, C. Garbe, H. Gogas, J. Schachter, G. Linette, et al. (2013). "Phase III randomized clinical trial comparing tremelimumab with standard-of-care chemotherapy in patients with advanced melanoma". In: *Journal of clinical oncology* 31.5, pp. 616–622 (cit. on p. 69).

Ristl, R., N. M. Ballarini, H. Götte, A. Schüler, M. Posch, and F. König (2021). "Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology". In: *Pharmaceutical statistics* 20.1, pp. 129–145 (cit. on p. 2).

Rizvi, N. A., B. C. Cho, N. Reinmuth, K. H. Lee, A. Luft, M.-J. Ahn, M. M. Van Den Heuvel, M. Cobo, D. Vicente, A. Smolin, et al. (2020). "Durvalumab with or without tremelimumab vs standard chemotherapy in first-line treatment of metastatic non–small cell lung cancer: the MYSTIC phase 3 randomized clinical trial". In: *JAMA oncology* 6.5, pp. 661–674 (cit. on p. 49).

Robert, C., G. V. Long, B. Brady, C. Dutriaux, A. M. Di Giacomo, L. Mortier, P. Rutkowski, J. C. Hassel, C. M. McNeil, E. A. Kalinka, et al. (2020). "Five-year outcomes with nivolumab in patients with wild-type BRAF advanced melanoma". In: *Journal of Clinical Oncology* 38.33, pp. 3937–3946 (cit. on p. 209).

Robert, C., G. V. Long, B. Brady, C. Dutriaux, M. Maio, L. Mortier, J. C. Hassel, P. Rutkowski, C. McNeil, E. Kalinka-Warzocha, et al. (2015). "Nivolumab in previously untreated melanoma without BRAF mutation". In: *New England journal of medicine* 372.4, pp. 320–330 (cit. on pp. 69, 209).

Robert, C., L. Thomas, I. Bondarenko, S. O'Day, J. Weber, C. Garbe, C. Lebbe, J.-F. Baurain, A. Testori, J.-J. Grob, et al. (2011). "Ipilimumab plus dacarbazine for previously untreated metastatic melanoma". In: *New England Journal of Medicine* 364.26, pp. 2517–2526 (cit. on p. 69).

Roychoudhury, S. and B. Neuenschwander (2020). "Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint". In: *Statistics in medicine* 39.7, pp. 984–995 (cit. on p. 55).

Royston, P. and M. K. Parmar (2013). "Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome". In: *BMC Medical Research Methodology* 13.1. DOI: 10.1186/1471-2288-13-152 (cit. on p. 52).

Rufibach, K., H. U. Burger, and M. Abt (2016). "Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development". In: *Pharmaceutical Statistics* 15.5, pp. 438–446. DOI: 10.1002/pst.1764 (cit. on pp. 10, 85).

Sabin, T., J. Matcham, S. Bray, A. Copas, and M. K. Parmar (2014). "A quantitative process for enhancing end of phase 2 decisions". In: *Statistics in Biopharmaceutical Research* 6.1, pp. 67–77 (cit. on p. 21).

Sackett, D. L., W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson (1996). *Evidence based medicine: what it is and what it isn't* (cit. on p. 1).

Saint-Hilary, G., V. Barboux, M. Pannaux, M. Gasparini, V. Robert, and G. Mastrantonio (2019). "Predictive probability of success using surrogate endpoints". In: *Statistics in Medicine* 38.10, pp. 1753–1774 (cit. on p. 10).

Salsbury, J. (2025). *DTEAssurance: Assurance Methods for Clinical Trials with a Delayed Treatment Effect*. R package version 1.0.1. DOI: 10.32614/CRAN.package.DTEAssurance (cit. on p. 5).

Salsbury, J. A., J. E. Oakley, S. A. Julious, and L. V. Hampson (2024). "Assurance methods for designing a clinical trial with a delayed treatment effect". en. In: *Statistics in Medicine* 43.19, pp. 3595–3612. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.10136 (cit. on p. 49).

SAS Institute Inc. (2014). *SAS/STAT® 9.4 User's Guide*. Version 9.4. SAS Institute Inc. Cary, NC (cit. on p. 100).

Saville, B. R., J. T. Connor, G. D. Ayers, and J. Alvarez (2014). "The utility of Bayesian predictive probabilities for interim monitoring of clinical trials". In: *Clinical Trials* 11.4, pp. 485–493. DOI: 10.1177/1740774514531352 (cit. on pp. 85, 108, 109).

Schmidli, H., S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander (2014). "Robust meta-analytic-predictive priors in clinical trials

with historical control information". In: *Biometrics* 70.4, pp. 1023–1032. DOI: 10.1111/biom.12242 (cit. on pp. 22, 55, 65, 198, 199).

Shitara, K. et al. (2020). "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". In: *JAMA oncology* 6.10, pp. 1571–1580. DOI: 10.1001/jamaoncol.2020.3370 (cit. on p. 49).

Slud, E. and L. Wei (1982). "Two-sample repeated significance tests based on the modified Wilcoxon statistic". In: *Journal of the American Statistical Association* 77.380, pp. 862–868 (cit. on p. 93).

Soares, M., A. Colson, L. Bojke, S. Ghabri, O. U. Garay, J. K. Felli, K. Lee, E. Molsen-David, O. Morales-Napoles, V. A. Shaffer, et al. (2024). "Recommendations on the use of structured expert elicitation protocols for healthcare decision making: a good practices report of an ISPOR task force". In: *Value in Health* 27.11, pp. 1469–1478 (cit. on p. 37).

Spiegelhalter, D. J. and L. S. Freedman (1986). "A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion". en. In: *Statistics in Medicine* 5.1, pp. 1–13. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.4780050103 (cit. on pp. 9, 20, 24).

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian approaches to clinical trials and health-care evaluation*. John Wiley & Sons (cit. on p. 11).

Spiegelhalter, D. J., L. S. Freedman, and P. R. Blackburn (1986). "Monitoring clinical trials: Conditional or predictive power?" In: *Controlled Clinical Trials* 7.1, pp. 8–17. DOI: 10.1016/0197-2456(86)90003-6 (cit. on pp. 9, 85, 91).

Spiegelhalter, D. J., L. S. Freedman, and M. K. B. Parmar (1994). "Bayesian Approaches to Randomized Trials". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 157.3, pp. 357–416. ISSN: 09641998, 1467985X (cit. on pp. 9, 20).

Spigel, D., F. De Marinis, G. Giaccone, N. Reinmuth, A. Vergnenegre, C. Barrios, M. Morise, E. Felip, Z. Andric, S. Geater, et al. (2019). "IMpower110: Interim overall survival (OS) analysis of a phase III study of atezolizumab (atezo) vs platinum-based chemotherapy (chemo) as first-line (1L) treatment (tx) in PD-L1–selected NSCLC". In: *Annals of Oncology* 30, p. v915 (cit. on p. 209).

Sprung, J., A. Basem, A. Gottlieb, C. Mayhew, J. Hammel, P. J. Levy, P. O'Hara, and N. R. Hertzer (2000). "Analysis of risk factors for myocardial infarction and cardiac

mortality after major vascular surgery". In: *Anesthesiology* 93(1), pp. 129–140. DOI: 10.1097/00000542-200007000-00023 (cit. on p. 17).

Sully, B. G., S. A. Julious, and J. Nicholl (2014). "An investigation of the impact of futility analysis in publicly funded trials". In: *Trials* 15, pp. 1–9 (cit. on pp. 103, 104).

Temple, J. R. and J. R. Robertson (2021). "Conditional assurance: the answer to the questions that should be asked within drug development". In: *Pharmaceutical Statistics* 20.6, pp. 1102–1111 (cit. on p. 20).

Thall, P. F. and R. Simon (1994). "A Bayesian approach to establishang sample size and monitoring criteria for phase II clinical trials". In: *Controlled Clinical Trials* 15.6, pp. 463–481. DOI: 10.1016/0197-2456(94)90004-3 (cit. on pp. 108, 109).

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4, pp. 285–294 (cit. on p. 81).

Todd, S., A. Whitehead, N. Stallard, and J. Whitehead (2001). "Interim analyses and sequential designs in phase III studies". In: *British journal of clinical pharmacology* 51.5, pp. 394–399 (cit. on p. 93).

Tsiatis, A. A. (1981). "The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time". In: *Biometrika* 68.1, pp. 311–315 (cit. on p. 82).

Tsiatis, A. A. (1982). "Repeated significance testing for a general class of statistics used in censored survival analysis". In: *Journal of the American Statistical Association* 77.380, pp. 855–861 (cit. on p. 82).

U.S. Food and Drug Administration (2019). *Adaptive Designs for Clinical Trials of Drugs and Biologics (Final Guidance)*. Guidance for Industry, finalized Nov 29 2019; available in Federal Register Dec 2 2019; Accessed Jul 14 2025 (cit. on pp. 2, 82, 102, 105, 110, 170).

Vandemeulebroecke, M. (2008). "Group sequential and adaptive designs–a review of basic concepts and points of discussion". In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50.4, pp. 541–557 (cit. on pp. 104, 105, 110).

Viele, K., S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. (2014). "Use of historical control data

for assessing treatment effects in clinical trials". In: *Pharmaceutical statistics* 13.1, pp. 41–54 (cit. on p. 22).

Walley, R. J., C. L. Smith, J. D. Gale, and P. Woodward (2015). "Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study". In: *Pharmaceutical Statistics* 14.3, pp. 205–215 (cit. on p. 21).

Walter, S., H. Han, G. Guyatt, D. Bassler, N. Bhatnagar, V. Gloy, S. Schandelmaier, and M. Briel (2020). "A systematic survey of randomised trials that stopped early for reasons of futility". In: *BMC medical research methodology* 20.1, p. 10 (cit. on p. 84).

Wang, S. K. and A. A. Tsiatis (1987). "Approximately optimal one-parameter boundaries for group sequential trials". In: *Biometrics*, pp. 193–199 (cit. on p. 95).

Wang, Y., H. Fu, P. Kulkarni, and C. Kaiser (2013). "Evaluating and utilizing probability of study success in clinical development". In: *Clinical Trials* 10.3, pp. 407–413. DOI: 10.1177/1740774513478229 (cit. on pp. 21, 85).

Wassmer, G. and F. Pahlke (2025). *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*. R package version 4.2.0. DOI: 10.32614/CRAN.package.rpact (cit. on p. 100).

Werner, C., T. Bedford, R. M. Cooke, A. M. Hanea, and O. Morales-Napoles (2017). "Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions". In: *European Journal of Operational Research* 258.3, pp. 801–819 (cit. on p. 31).

Werner, C., A. M. Hanea, and O. Morales-Nápoles (2017). "Eliciting multivariate uncertainty from experts: considerations and approaches along the expert judgement process". In: *Elicitation: The science and art of structuring judgement*. Springer, pp. 171–210 (cit. on p. 31).

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. John Wiley & Sons (cit. on pp. 93, 104).

Wieand, S., G. Schroeder, and J. R. O'Fallon (1994). "Stopping when the experimental regimen does not appear to help". In: *Statistics in Medicine* 13.13-14, pp. 1453–1458. DOI: 10.1002/sim.4780131321 (cit. on p. 113).

Williams, C. J., K. J. Wilson, and N. Wilson (2021). "A comparison of prior elicitation aggregation using the classical method and SHELF". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 184.3, pp. 920–940 (cit. on p. 35).

Williamson, S. F., S. V. Tishkovskaya, and K. J. Wilson (2025). "Hybrid sample size calculations for cluster randomised trials using assurance". In: *Clinical Trials*, p. 17407745241312635 (cit. on p. 20).

Wilson, K. J. (2017). "An investigation of dependence in expert judgement studies with multiple experts". In: *International Journal of Forecasting* 33.1, pp. 325–336 (cit. on p. 32).

Wilson, K. J. (2023). "Bayesian design and analysis of two-arm cluster randomized trials using assurance". In: *Statistics in Medicine* 42.25, pp. 4517–4531 (cit. on p. 20).

Wolchok, J. D., V. Chiarion-Sileni, R. Gonzalez, P. Rutkowski, J.-J. Grob, C. L. Cowey, C. D. Lao, J. Wagstaff, D. Schadendorf, P. F. Ferrucci, et al. (2017). "Overall survival with combined nivolumab and ipilimumab in advanced melanoma". In: *New England Journal of Medicine* 377.14, pp. 1345–1356 (cit. on p. 209).

Wolchok, J. D., V. Chiarion-Sileni, P. Rutkowski, C. L. Cowey, D. Schadendorf, J. Wagstaff, P. Queirolo, R. Dummer, M. O. Butler, A. G. Hill, et al. (2024). "Final, 10-year outcomes with nivolumab plus ipilimumab in advanced melanoma". In: *The New England journal of medicine* 392.1, p. 11 (cit. on p. 209).

Wu, J., Y. Li, and L. Zhu (2023). "Group sequential designs for cancer immunotherapy trial with delayed treatment effect". In: *Journal of Biopharmaceutical Statistics* 34.1, pp. 1–15. DOI: 10.1080/10543406.2023.2170403 (cit. on pp. 112, 130).

Yen, C.-J., N. Kiyota, N. Hanai, S. Takahashi, T. Yokota, S. Iwae, Y. Shimizu, R.-L. Hong, M. Goto, J.-H. Kang, W. S. K. Li, R. L. Ferris, M. Gillison, T. Endo, V. Jayaprakash, and M. Tahara (2020). "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)". In: *Head & Neck* 42.10, pp. 2852–2862. DOI: 10.1002/hed.26331 (cit. on pp. xiv, xv, 51, 73, 209).

Yi, J., L. Fang, and Z. Su (2012). "Hybridization of conditional and predictive power for futility assessment in sequential clinical trials with time-to-event outcomes: a resampling approach". In: *Contemporary Clinical Trials* 33.1, pp. 138–142 (cit. on pp. 108, 109).

Zapata-Vázquez, R. E., A. O'Hagan, and L. S. B. and (2014). "Eliciting expert judgements about a set of proportions". In: *Journal of Applied Statistics* 41.9, pp. 1919–1933. DOI: 10.1080/02664763.2014.898131. eprint: https://doi.org/10.1080/02664763.2014.898131 (cit. on p. 31).

Appendices

Appendix A

Derivations

A.1 Derivation of the Critical Value for the Hypothesis Test

Consider a two-arm clinical trial comparing a treatment group to a control group, where the primary endpoint is continuous. Let $\hat{\delta}$ denote the observed difference in sample means between the treatment and control arms, and assume the following model:

$$\hat{\delta} \sim \mathcal{N}\left(\delta, \frac{2\sigma^2}{n}\right),$$

where δ is the true treatment effect, σ^2 is the common population variance, and n is the number of patients allocated to each arm (equal allocation assumed).

The goal is to test the one-sided hypothesis:

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta > 0,$$

at a significance level α.

Under the null hypothesis H_0, the distribution of $\hat{\delta}$ is

$$\hat{\delta} \sim \mathcal{N}\left(0, \frac{2\sigma^2}{n}\right).$$

The standardised test statistic is then

$$Z = \frac{\hat{\delta} - 0}{\sqrt{2\sigma^2/n}} = \frac{\hat{\delta}}{\sqrt{2\sigma^2/n}}.$$

For a one-sided test at level α, the null hypothesis is rejected if the test statistic exceeds the $(1-\alpha)$-quantile of the standard normal distribution:

$$Z > Z_{1-\alpha}.$$

Rearranging the inequality for $\hat{\delta}$, we have

$$\frac{\hat{\delta}}{\sqrt{2\sigma^2/n}} > Z_{1-\alpha} \implies \hat{\delta} > \sqrt{\frac{2\sigma^2}{n}} Z_{1-\alpha} = \frac{\sqrt{2}\sigma}{\sqrt{n}} Z_{1-\alpha}.$$

A.2 MAP Prior for Historical Control Data - Exponential

In this section, we describe how historical control data can be incorporated into the design of a future survival trial using a Bayesian hierarchical framework. Specifically, we employ the Meta-Analytic-Predictive (MAP) prior approach, originally introduced by Neuenschwander et al., 2010; Schmidli et al., 2014 and later extended to time-to-event data by Bertsche et al., 2019. This method enables the combination of information from H historical control datasets to form a single predictive distribution for the control group hazard rate in the upcoming trial.

For each historical study $h = 1, \ldots, H$, we observe n_h event or censoring times t_{ih}, among which d_h are observed events and $n_h - d_h$ are censored observations. Assuming that event times follow an exponential distribution,

$$T_{ih} \sim \text{Exp}(\lambda_h),$$

and considering a Type II censoring scheme, the sum of observed survival times

$$T_{+h} = \sum_{i=1}^{d_h} t_{ih}$$

follows a Gamma distribution,

$$T_{+h} \sim \text{Gamma}(d_h, \lambda_h). \tag{A.1}$$

We assume the same distributional form holds for the future control group, denoted C, yielding

$$T_{+C} \sim \text{Gamma}(d_C, \lambda_C).$$

To capture between-study heterogeneity, we model the log hazard rates of the historical studies and the future control group as exchangeable and normally distributed:

$$\log(\lambda_1), \ldots, \log(\lambda_H), \log(\lambda_C) \sim \mathcal{N}(\mu, \tau^2). \tag{A.2}$$

Here, μ represents the overall mean log hazard rate, and τ captures the between-study variability.

Using the historical data, the MAP prior for the future control group hazard rate λ_C is defined as the posterior predictive distribution

$$\pi(\lambda_C \mid T_{+1}, \ldots, T_{+H}, \mu, \tau) =: \pi(\lambda_C). \tag{A.3}$$

A full Bayesian model requires specifying priors for the hyperparameters μ and τ. Following Schmidli et al., 2014, we assign a weakly informative normal prior to μ, with mean $\mu_0 = 0$ and variance $\sigma_0^2 = 1000$, and a half-normal prior on the standard deviation τ. Given the critical role of τ in governing borrowing strength, sensitivity analyses are recommended, particularly when the number of historical studies is small.

This hierarchical modelling approach allows us to formally quantify uncertainty and heterogeneity across historical datasets, improving the robustness and informativeness of the prior for the control group hazard in the future trial.

A.3 Log-likelihood expressions for delayed treatment effect models

In this appendix we provide the log-likelihood expressions for the exponential and Weibull delayed treatment effect models used in Section 6.4. Recall the index sets

$$\mathcal{C} = \{i : z_i = 0\}, \qquad \mathcal{T}_{\le T} = \{i : z_i = 1,\ x_i \le T\}, \qquad \mathcal{T}_{>T} = \{i : z_i = 1,\ x_i > T\},$$

and let $\mathcal{D}$ denote the full set of observed times and event indicators $\{(x_i, y_i)\}$.

A.3.1 Exponential Model

For the exponential model, the parameter vector is $\boldsymbol{\theta}_E = (\lambda_c, \mathrm{HR}^*, T)$. The individual likelihood contributions are

$$L_i = \begin{cases} \lambda_c^{y_i} \exp(-\lambda_c x_i), & i \in \mathcal{C} \cup \mathcal{T}_{\le T}, \\ (\lambda_c \mathrm{HR}^*)^{y_i} \exp\{-\lambda_c T - \lambda_c \mathrm{HR}^* (x_i - T)\}, & i \in \mathcal{T}_{>T}, \end{cases}$$

and the log-likelihood is

$$\ell_E(\boldsymbol{\theta}_E \mid \mathcal{D}) = \log \mathcal{L}(\boldsymbol{\theta}_E \mid \mathcal{D}) \\ = \sum_{i \in \mathcal{C} \cup \mathcal{T}_{\leq T}} \big[y_i \log \lambda_c - \lambda_c x_i \big] + \sum_{i \in \mathcal{T}_{>T}} \big[y_i (\log \lambda_c + \log \mathrm{HR}^*) - \lambda_c T - \lambda_c \mathrm{HR}^* (x_i - T) \big]. \tag{A.4}$$

A.3.2 Weibull Model

For the Weibull model, the parameter vector is $\boldsymbol{\theta}_W = (\lambda_c, \gamma_c, \mathrm{HR}^*, T)$. The control-arm survival is $S_c(t) = \exp\{-(\lambda_c t)^{\gamma_c}\}$, with cumulative hazard $H_C(t) = (\lambda_c t)^{\gamma_c}$ and hazard $h_C(t) = \gamma_c \lambda_c^{\gamma_c} t^{\gamma_c - 1}$.

The individual likelihood contributions can be written as

$$L_i = \begin{cases} \big(\gamma_c \lambda_c^{\gamma_c} x_i^{\gamma_c - 1}\big)^{y_i} \exp(-(\lambda_c x_i)^{\gamma_c}), & i \in \mathcal{C} \cup \mathcal{T}_{\leq T}, \\ \big(\mathrm{HR}^* \gamma_c \lambda_c^{\gamma_c} x_i^{\gamma_c - 1}\big)^{y_i} \exp\big\{-(\lambda_c T)^{\gamma_c} - \mathrm{HR}^* \big((\lambda_c x_i)^{\gamma_c} - (\lambda_c T)^{\gamma_c}\big)\big\}, & i \in \mathcal{T}_{>T}. \end{cases}$$

The corresponding log-likelihood is

$$\ell_W(\boldsymbol{\theta}_W \mid \mathcal{D}) = \log \mathcal{L}(\boldsymbol{\theta}_W \mid \mathcal{D}) \\ = \sum_{i \in \mathcal{C} \cup \mathcal{T}_{\leq T}} \big[y_i \{ \log \gamma_c + \gamma_c \log \lambda_c + (\gamma_c - 1) \log x_i \} - (\lambda_c x_i)^{\gamma_c} \big] \\ + \sum_{i \in \mathcal{T}_{>T}} \big[y_i \{ \log \mathrm{HR}^* + \log \gamma_c + \gamma_c \log \lambda_c + (\gamma_c - 1) \log x_i \} \\ - (\lambda_c T)^{\gamma_c} - \mathrm{HR}^* \big((\lambda_c x_i)^{\gamma_c} - (\lambda_c T)^{\gamma_c} \big) \big]. \tag{A.5}$$

A.4 Simulation Formulae for Predictive Probability Under Delayed Treatment Effects

This appendix provides explicit simulation formulae for each risk-set category described in Section 6.4.3, for both the exponential and Weibull delayed-effect models. In all cases, administrative censoring is enforced by the trial calendar: simulated event times that exceed the calendar time of the final analysis are censored at that point. Let x_i denote the accumulated follow-up time at the interim analysis for patient i, and let $\boldsymbol{\theta}^{(m)} = (\lambda_c^{(m)}, \mathrm{HR}^{*(m)}, \tau^{(m)})$ (exponential) or $\boldsymbol{\theta}^{(m)} = (\lambda_c^{(m)}, \gamma_c^{(m)}, \mathrm{HR}^{*(m)}, \tau^{(m)})$ (Weibull) denote the posterior draw at iteration m.

A.4.1 Exponential Model

Under the exponential model, $\lambda_e = \lambda_c \cdot \text{HR}^*$.

- **Patients not yet enrolled — control arm:** event times are simulated from the full exponential model:
$$\tilde{t}_i \sim \text{Exp}(\lambda_c).$$

- **Patients not yet enrolled — treatment arm:** event times are simulated using inverse CDF sampling from the piecewise hazard. Let $u \sim \text{Uniform}(0, 1)$ and $C_\tau = \exp(-\lambda_c \tau)$ denote the survival probability at the delay time. Then:
$$\tilde{t}_i = \begin{cases} -\dfrac{\log(u)}{\lambda_c} & \text{if } u > C_\tau, \\ \dfrac{1}{\lambda_e}\left(\lambda_e \tau - \log(u) - \lambda_c \tau\right) & \text{otherwise.} \end{cases}$$

- **Censored control patients:** using the memoryless property of the exponential distribution, the residual event time is simulated as:
$$\tilde{t}_i \sim x_i + \text{Exp}(\lambda_c).$$

- **Censored treated patients in the pre-delay region ($x_i < \tau$):** let $u \sim \text{Uniform}(0, 1)$ and $p = 1 - \exp(-\lambda_c(\tau - x_i))$ denote the probability of the event occurring before τ conditional on survival to x_i. Then:
$$\tilde{t}_i = \begin{cases} x_i + \dfrac{-\log(1 - u \cdot p)}{\lambda_c} & \text{if } u \leq p, \\ \tau + \dfrac{-\log(u) - \lambda_c(\tau - x_i)}{\lambda_e} & \text{otherwise.} \end{cases}$$

- **Censored treated patients in the post-delay region ($x_i > \tau$):** using the memoryless property, the residual event time is simulated as:
$$\tilde{t}_i \sim x_i + \text{Exp}(\lambda_e).$$

- **Patients who have already experienced the event:** no further simulation is required.

A.4.2 Weibull Model

Under the Weibull model, the shape parameter γ_c is assumed equal across both arms after the delay, and $\lambda_e = \lambda_c \cdot \mathrm{HR}^{*1/\gamma_c}$. Let $H(t) = (\lambda_c t)^{\gamma_c}$ denote the cumulative hazard under the control arm.

- **Patients not yet enrolled — control arm:** let $u \sim \mathrm{Uniform}(0,1)$. Event times are simulated using inverse CDF sampling:

$$\tilde{t}_i = \frac{(-\log u)^{1/\gamma_c}}{\lambda_c}.$$

- **Patients not yet enrolled — treatment arm:** let $u \sim \mathrm{Uniform}(0,1)$ and $C_\tau = \exp\{-(\lambda_c \tau)^{\gamma_c}\}$. Then:

$$\tilde{t}_i = \begin{cases} \dfrac{(-\log u)^{1/\gamma_c}}{\lambda_c} & \text{if } u > C_\tau, \\ \left(\dfrac{-\log u - (1-\mathrm{HR}^*)(\lambda_c\tau)^{\gamma_c}}{\mathrm{HR}^*}\right)^{1/\gamma_c} \dfrac{1}{\lambda_c} & \text{otherwise.} \end{cases}$$

- **Censored control patients:** let $v \sim \mathrm{Uniform}(0,1)$. The residual event time is simulated using conditional inverse CDF sampling:

$$\tilde{t}_i = \frac{(H(x_i) - \log v)^{1/\gamma_c}}{\lambda_c}.$$

- **Censored treated patients in the pre-delay region ($x_i < \tau$):** let $H_\tau = (\lambda_c\tau)^{\gamma_c}$, $H_{x_i} = (\lambda_c x_i)^{\gamma_c}$, $S_{x_i} = \exp(-H_{x_i})$, and $S_\tau = \exp(-H_\tau)$. The probability of the event occurring before τ conditional on survival to x_i is:

$$p = 1 - \exp(-(H_\tau - H_{x_i})).$$

  Let $u \sim \mathrm{Uniform}(0,1)$. Then:

  - *Early branch* ($u \le p$): the event occurs before τ. Let $v \sim \mathrm{Uniform}(0,1)$:

$$\tilde{t}_i = \frac{(-\log(S_{x_i} - v(S_{x_i} - S_\tau)))^{1/\gamma_c}}{\lambda_c}.$$

  - *Late branch* ($u > p$): the event occurs after τ. Let $v \sim \mathrm{Uniform}(0,1)$:

$$\tilde{t}_i = \frac{\left(H_\tau - \dfrac{\log v}{\mathrm{HR}^*}\right)^{1/\gamma_c}}{\lambda_c}.$$

- **Censored treated patients in the post-delay region ($x_i > \tau$):** let $v \sim \text{Uniform}(0, 1)$ and $H^t_{x_i} = (\lambda_e x_i)^{\gamma_c}$. The residual event time is simulated as:

$$\tilde{t}_i = \frac{\left(H^t_{x_i} - \log v\right)^{1/\gamma_c}}{\lambda_e}.$$

- **Patients who have already experienced the event:** no further simulation is required.

Appendix B

JAGS Model Specifications

This appendix provides the JAGS model code used for Bayesian posterior inference at interim analyses, as described in Section 6.4.2. Separate models are given for the exponential and Weibull specifications. In both cases, the Poisson zeros trick is used to specify a custom likelihood (Ntzoufras, 2011): each observation contributes a term l_i to the log-likelihood, and the zeros trick encodes this as a Poisson observation with mean $C - l_i$, where $C = 10{,}000$ is a large constant ensuring non-negative means.

The data are ordered so that control patients occupy indices $1, \ldots, n$ and treatment patients occupy indices $n+1, \ldots, m$. The categorical variable Z governs the treatment effect structure:

- $Z = 1$: no treatment effect (HR $= 1$, no delay);
- $Z = 2$: immediate treatment effect (HR $=$ HR*, no delay);
- $Z = 3$: delayed treatment effect (HR $=$ HR*, delay $= \tau$).

The prior probabilities for Z are specified as $\pi = (1 - P_{\mathcal{S}},\ P_{\mathcal{S}}(1 - P_{\text{DTE}}),\ P_{\mathcal{S}} \cdot P_{\text{DTE}})$, corresponding to the elicited probabilities of no separation, immediate separation, and delayed separation respectively.

B.1 Exponential Model

```
data {
  for (j in 1:m) {
```

```
    zeros[j] <- 0
  }
}
model {
  C <- 10000
  # Control arm likelihood (indices 1 to n)
  for (i in 1:n) {
    zeros[i] ~ dpois(zeros.mean[i])
    zeros.mean[i] <- -l[i] + C
    l[i] <- ifelse(data_event[i] == 1,
               log(lambda_c) - (lambda_c * data_time[i]),
               -(lambda_c * data_time[i]))
  }
  # Treatment arm likelihood (indices n+1 to m)
  for (i in (n+1):m) {
    zeros[i] ~ dpois(zeros.mean[i])
    zeros.mean[i] <- -l[i] + C
    l[i] <- ifelse(data_event[i] == 1,
               ifelse(data_time[i] < delay_time,
                 log(lambda_c) - (lambda_c * data_time[i]),
                 log(lambda_e) - lambda_e * (data_time[i] - delay_time)
                   - (delay_time * lambda_c)),
               ifelse(data_time[i] < delay_time,
                 -(lambda_c * data_time[i]),
                 -(lambda_c * delay_time)
                   - lambda_e * (data_time[i] - delay_time)))
  }
  # Treatment effect structure
  Z ~ dcat(pi[])
  HR_slab    ~ dgamma(c, d)       # prior for HR* (example)
  delay_slab ~ dgamma(a, b)       # prior for delay time (example)
  HR <- equals(Z, 1) * 1
        + (1 - equals(Z, 1)) * HR_slab
  delay_time <- equals(Z, 3) * delay_slab
  # Control hazard prior (example: exponential model)
  lambda_c <- -log(s1) / t1      # derived from elicited s1
  # Treatment hazard
  lambda_e <- lambda_c * HR
}
```

B.2 Weibull Model

```
data {
  for (j in 1:m) {
    zeros[j] <- 0
  }
}
model {
  C <- 10000
  # Control arm likelihood (indices 1 to n)
  for (i in 1:n) {
    zeros[i] ~ dpois(zeros.mean[i])
    zeros.mean[i] <- -l[i] + C
    l[i] <- ifelse(data_event[i] == 1,
                log(gamma_c) + gamma_c * log(lambda_c * data_time[i])
                  - (lambda_c * data_time[i])^gamma_c - log(data_time[i]),
                -(lambda_c * data_time[i])^gamma_c)
  }
  # Treatment arm likelihood (indices n+1 to m)
  for (i in (n+1):m) {
    zeros[i] ~ dpois(zeros.mean[i])
    zeros.mean[i] <- -l[i] + C
    l[i] <- ifelse(data_event[i] == 1,
                ifelse(data_time[i] < delay_time,
                  log(gamma_c) + gamma_c * log(lambda_c * data_time[i])
                    - (lambda_c * data_time[i])^gamma_c - log(data_time[i]),
                  log(gamma_c) + gamma_c * log(lambda_e)
                    + (gamma_c - 1) * log(data_time[i])
                    - lambda_e^gamma_c * (data_time[i]^gamma_c
                    - delay_time^gamma_c)
                    - (delay_time * lambda_c)^gamma_c),
                ifelse(data_time[i] < delay_time,
                  -(lambda_c * data_time[i])^gamma_c,
                  -(lambda_c * delay_time)^gamma_c
                    - lambda_e^gamma_c * (data_time[i]^gamma_c
                    - delay_time^gamma_c)))
  }
  # Treatment effect structure
  Z ~ dcat(pi[])
  HR_slab    ~ dgamma(c, d)       # prior for HR* (example)
  delay_slab ~ dgamma(a, b)       # prior for delay time (example)
  HR <- equals(Z, 1) * 1
```

```
        + (1 - equals(Z, 1)) * HR_slab
  delay_time <- equals(Z, 3) * delay_slab
  # Control parameters (example: Weibull model)
  s1 ~ dbeta(a_s1, b_s1)         # elicited prior for S(t1)
  delta ~ dbeta(a_delta, b_delta) # elicited prior for S(t1) - S(t2)
  gamma_c <- log(log(s1) / log(s1 - delta)) / log(t1 / t2)
  lambda_c <- pow(-log(s1), 1 / gamma_c) / t1
  # Treatment hazard
  lambda_e <- lambda_c * pow(HR, 1 / gamma_c)
}
```

Appendix C

Vivli Work - Case Studies

C.1 Publications

Table C.1: *Publications associated with each data set.*

Trial (NCT ID)	Main Publication	Updated / Interim Publication(s)
Trial A (NCT02409342)	Herbst, Giaccone, et al., 2020	Spigel et al., 2019, Jassem et al., 2021
Trial B (NCT01668784)	Motzer, Escudier, McDermott, et al., 2015	Motzer, Escudier, George, et al., 2020
Trial C (NCT01721772)	Robert, Long, Brady, Dutriaux, Maio, et al., 2015	Robert, Long, Brady, Dutriaux, Di Giacomo, et al., 2020
Trial D (NCT01844505)	Wolchok, Chiarion-Sileni, Gonzalez, et al., 2017	Wolchok, Chiarion-Sileni, Rutkowski, et al., 2024
Trial E (NCT01642004)	Brahmer, Reckamp, et al., 2015	Borghaei, Gettinger, et al., 2021
Trial F (NCT02477826)	Hellmann et al., 2019	Brahmer, Lee, et al., 2023
Trial G (NCT02105636)	Ferris et al., 2016	Yen et al., 2020
Trial H (NCT01673867)	Borghaei, Paz-Ares, et al., 2015	–

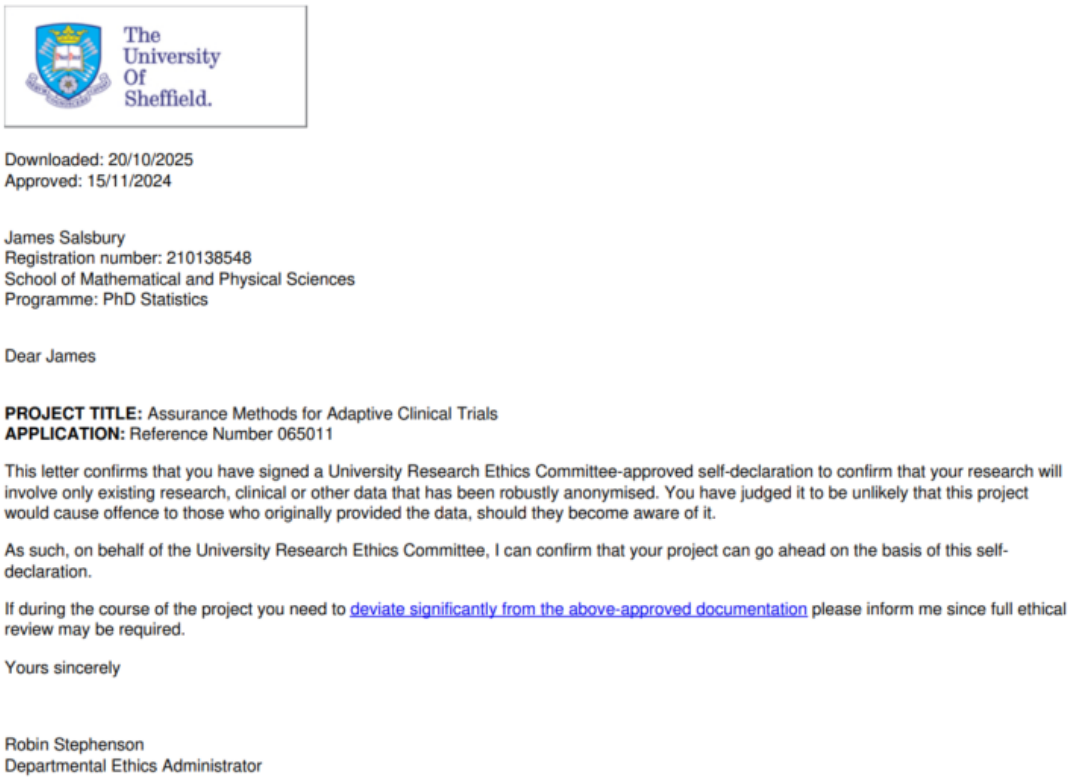
C.2 Vivli Data Request

A formal data request was submitted to the *Vivli* clinical trial data sharing platform for the project: *"Developing methodology for a delayed treatment effect present in a survival trial"* found here.

The application outlined the research objectives, proposed analyses, and justification for accessing individual participant data from seven completed survival trials. The complete six-page application is held on file with the author and is available upon request.

This research was reviewed and approved by the Departmental Ethics Administrator under reference number 065011. The study was classified as secondary analysis of anonymised clinical trial data and therefore posed no risk to participants. A copy of

the formal approval notice is shown in Figure C.1 for reference.



The University Of Sheffield.

Downloaded: 20/10/2025
Approved: 15/11/2024

James Salsbury
Registration number: 210138548
School of Mathematical and Physical Sciences
Programme: PhD Statistics

Dear James

PROJECT TITLE: Assurance Methods for Adaptive Clinical Trials
APPLICATION: Reference Number 065011

This letter confirms that you have signed a University Research Ethics Committee-approved self-declaration to confirm that your research will involve only existing research, clinical or other data that has been robustly anonymised. You have judged it to be unlikely that this project would cause offence to those who originally provided the data, should they become aware of it.

As such, on behalf of the University Research Ethics Committee, I can confirm that your project can go ahead on the basis of this self-declaration.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since full ethical review may be required.

Yours sincerely

Robin Stephenson
Departmental Ethics Administrator

Figure C.1: *Departmental ethical approval form for the Vivli case studies.*

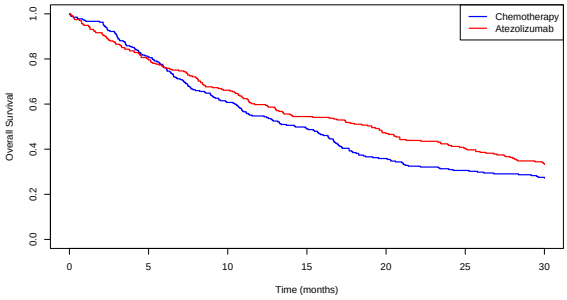
C.3 Supplementary Figures for All Trials

C.4 Supplementary Tables for All Trials

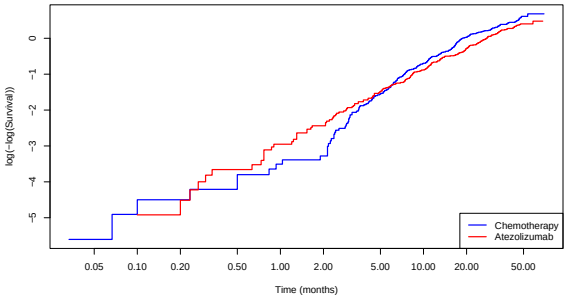
Table C.2: *Parameter estimates for progression-free survival (PFS) across trials under exponential and Weibull frameworks. τ denotes the estimated delay time (months); λ_C and γ_C are the scale and shape parameters for the control arm; HR* is the post-delay hazard ratio from the best-fitting model. The lowest AIC within each trial is shown in* ***bold****.*

Trial	Model	τ (mo)	λ_C	γ_C	HR*	AIC
B	Exponential	–	0.1149	–	0.8379	4751.2
	Weibull	–	0.1199	0.881	0.8645	4731.9
	Piecewise exponential	11.6	0.1232	–	0.2894	**4650.5**
	Piecewise Weibull	11.6	0.1233	0.994	0.2922	4652.5
C	Exponential	–	0.1910	–	0.2153	2207.9
	Weibull	–	0.2279	0.710	0.3014	2131.3
	Piecewise exponential	5.2	0.1668	–	0.0819	2048.5
	Piecewise Weibull	5.2	0.1680	0.924	0.0959	**2046.7**
E	Exponential	–	0.2061	–	0.5722	1229.3
	Weibull	–	0.2073	0.959	0.5881	1230.6
	Piecewise exponential	2.3	0.2149	–	0.3276	1193.3
	Piecewise Weibull	2.3	0.2173	1.177	0.2475	**1186.7**
G	Exponential	–	0.2134	–	0.8169	1618.2
	Weibull	–	0.2196	0.867	0.8664	1607.5
	Piecewise exponential	5.0	0.2536	–	0.2014	1521.6
	Piecewise Weibull	5.0	0.2512	1.130	0.1546	**1516.3**
H	Exponential	–	0.1555	–	0.7788	2966.9
	Weibull	–	0.1601	0.863	0.8334	2948.9
	Piecewise exponential	6.5	0.1757	–	0.2209	**2844.6**
	Piecewise Weibull	6.5	0.1747	1.047	0.2029	2845.1

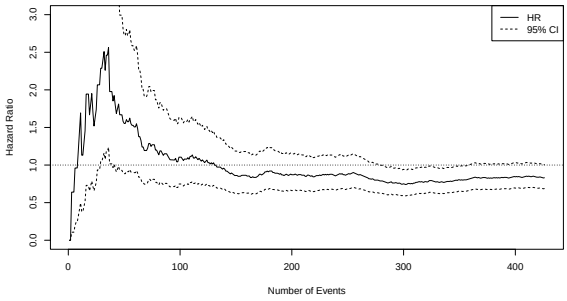
Trial A – Overall Survival (OS)



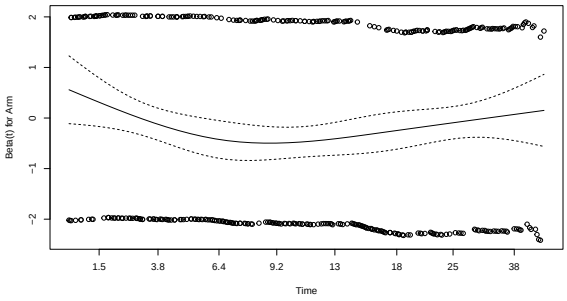
(a) *KM Plot*



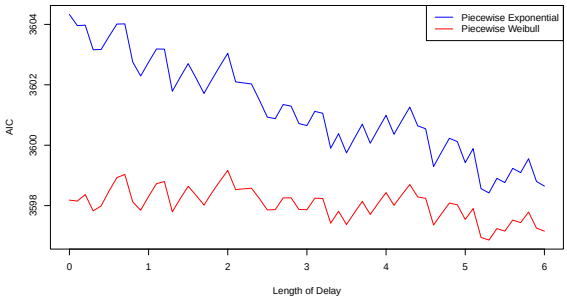
(b) *C-loglog Plot*



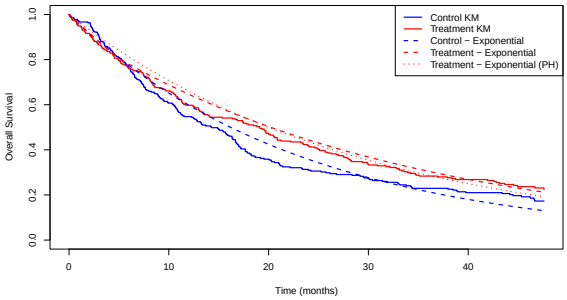
(c) *Hazard Ratio vs Events*



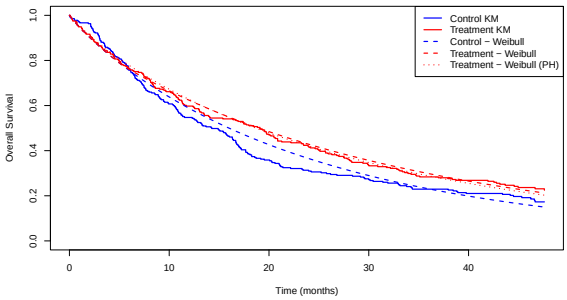
(d) *Schoenfeld residuals*



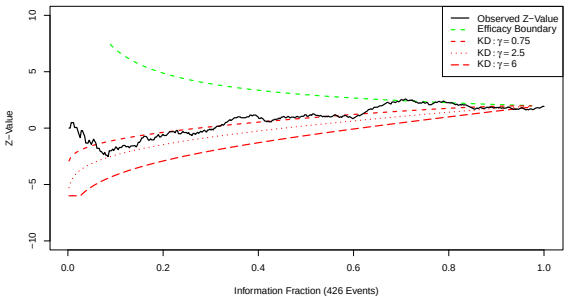
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

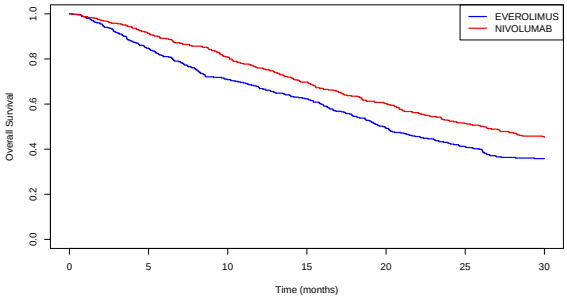


(g) *Piecewise Weibull Fit*

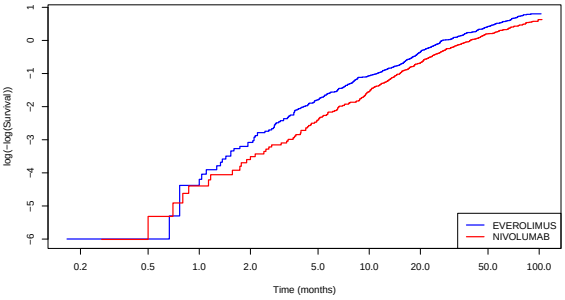


(h) *GSD Boundary Plot*

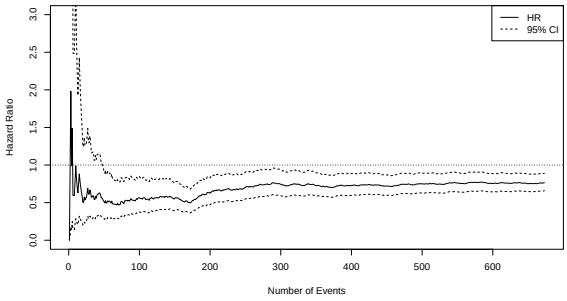
Trial B – Overall Survival (OS)



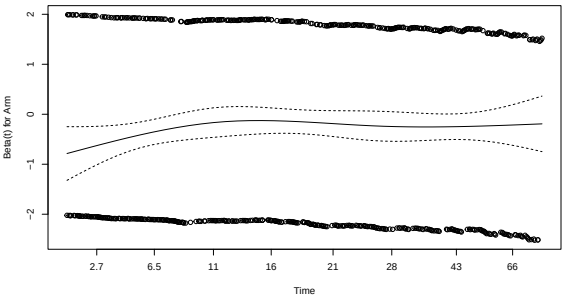
(a) *KM Plot*



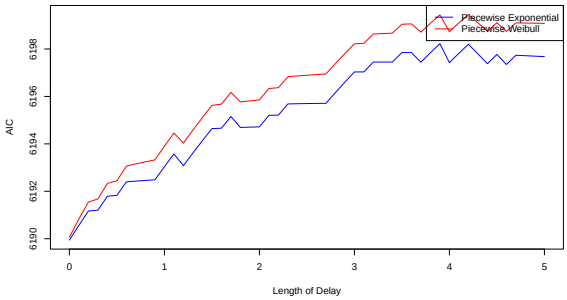
(b) *C-loglog*



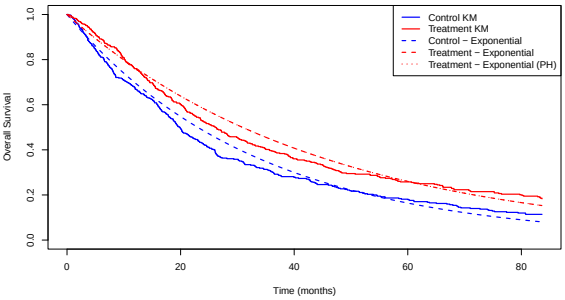
(c) *HR vs Events*



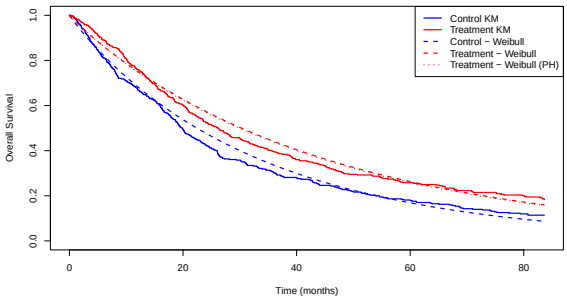
(d) *Schoenfeld residuals*



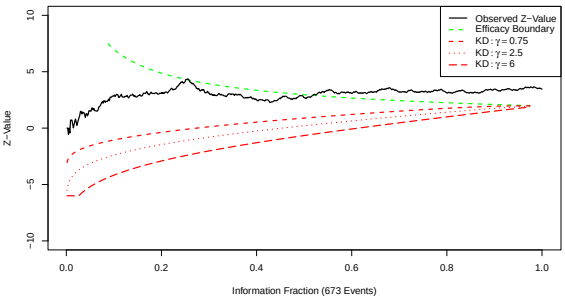
(e) *AIC vs Delay*



(f) *PWE Fit*

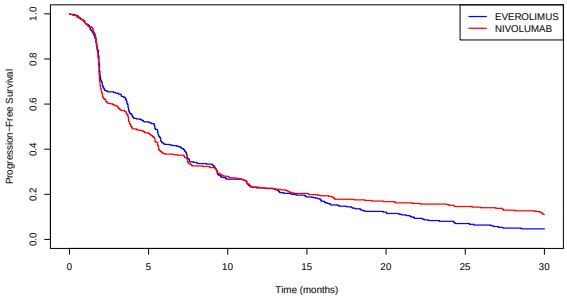


(g) *PWW Fit*

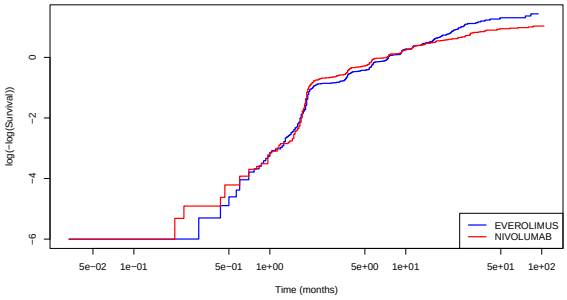


(h) *GSD Boundaries*

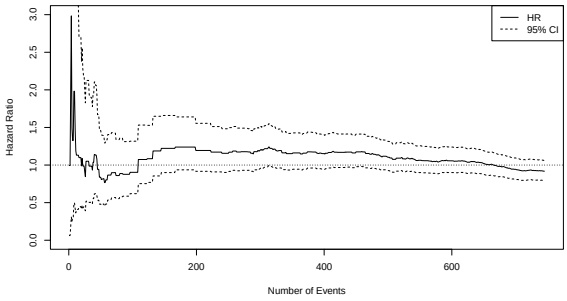
Trial B – Progression-Free Survival (PFS)



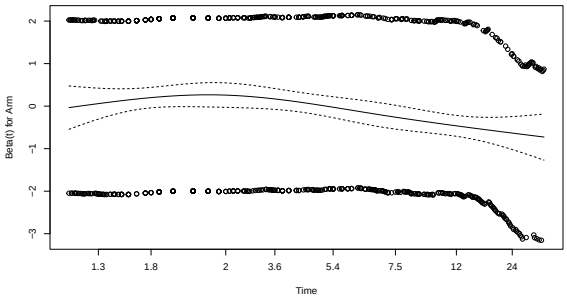
(a) *KM Plot*



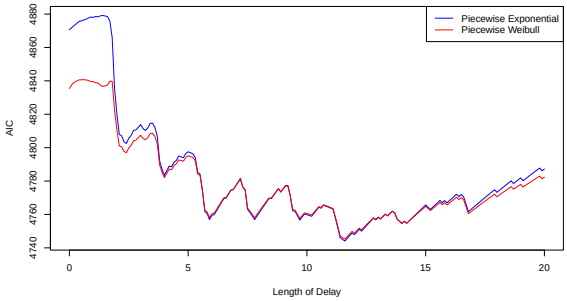
(b) *C-loglog Plot*



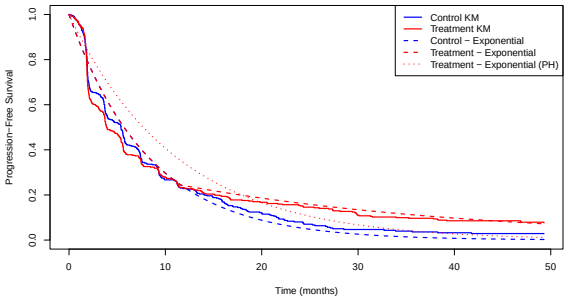
(c) *Hazard Ratio vs Events*



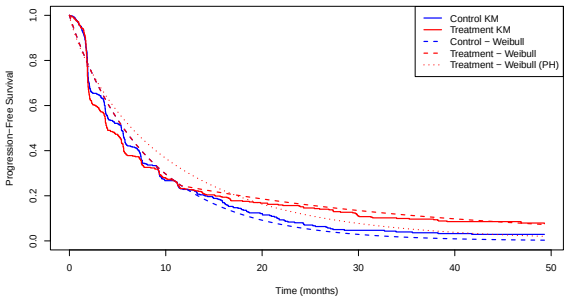
(d) *Schoenfeld residuals*



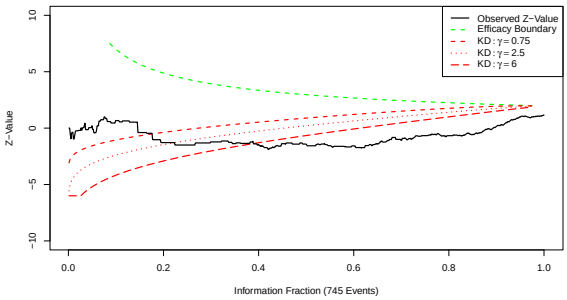
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

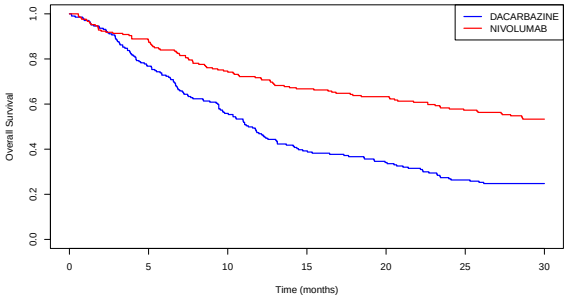


(g) *Piecewise Weibull Fit*

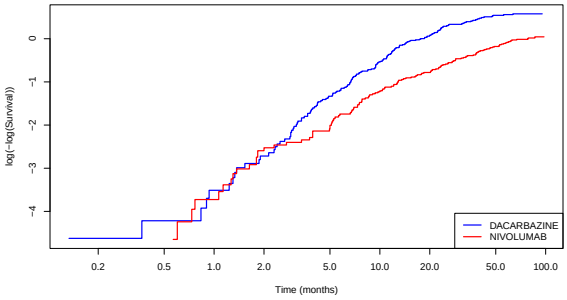


(h) *GSD Boundary Plot*

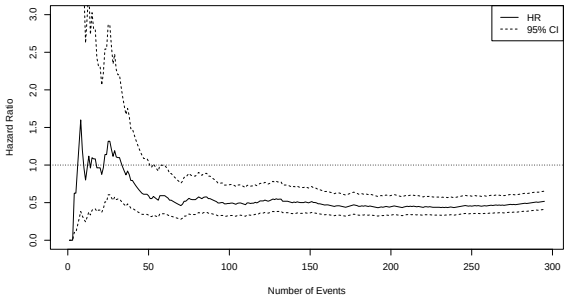
Trial C – Overall Survival (OS)



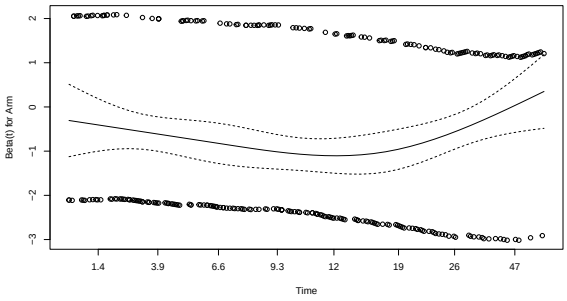
(a) *KM Plot*



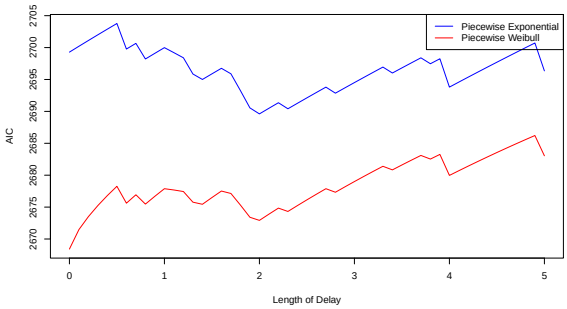
(b) *C-loglog Plot*



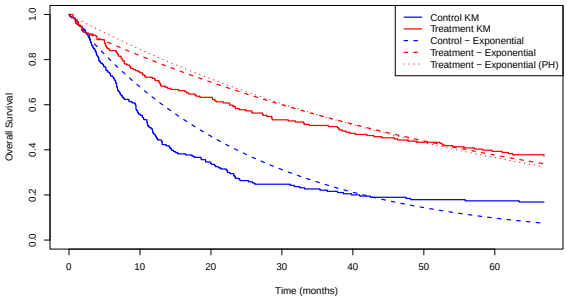
(c) *Hazard Ratio vs Events*



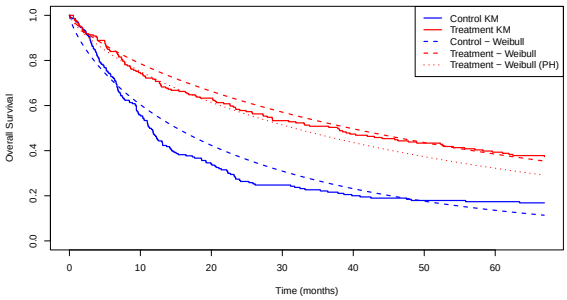
(d) *Schoenfeld residuals*



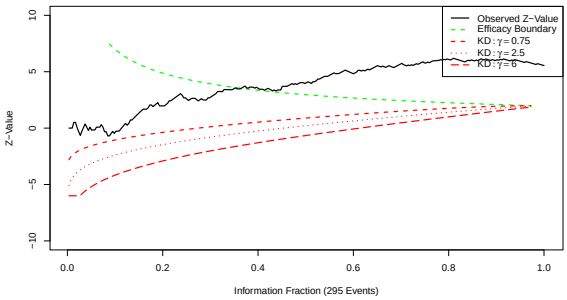
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

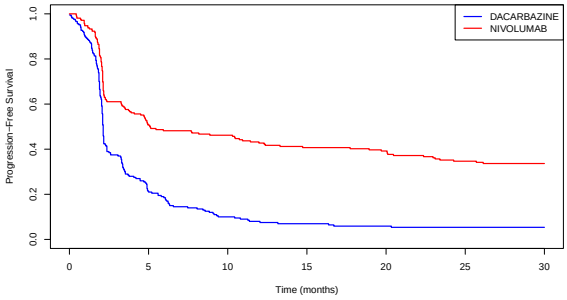


(g) *Piecewise Weibull Fit*

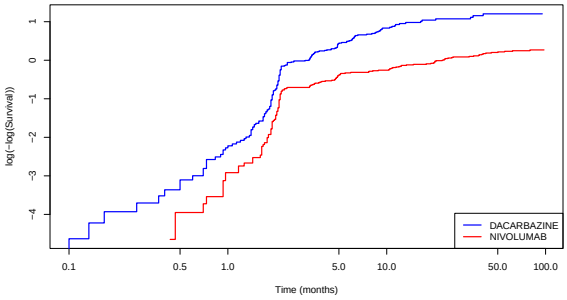


(h) *GSD Boundary Plot*

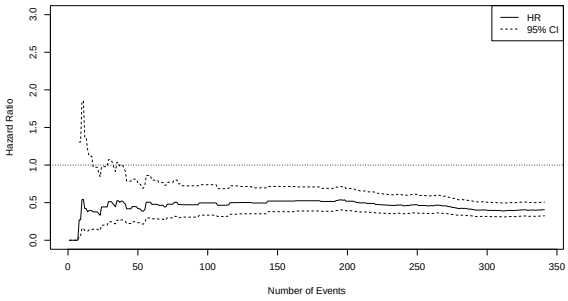
Trial C – Progression-Free Survival (PFS)



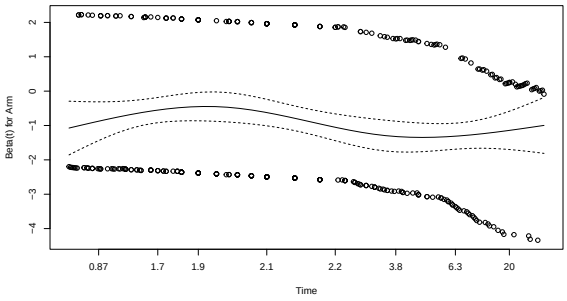
(a) *KM Plot*



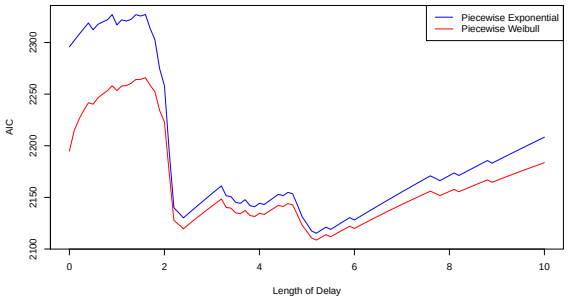
(b) *C-loglog Plot*



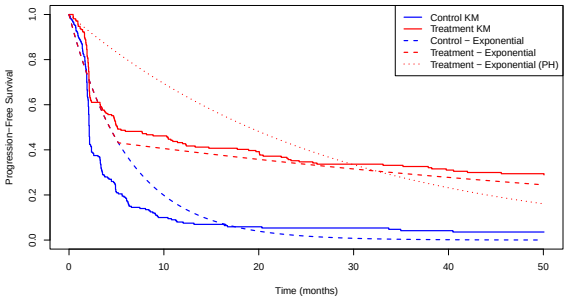
(c) *Hazard Ratio vs Events*



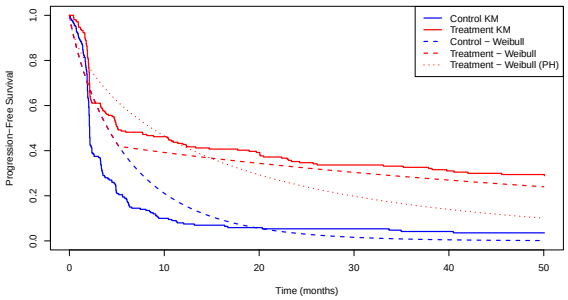
(d) *Schoenfeld residuals*



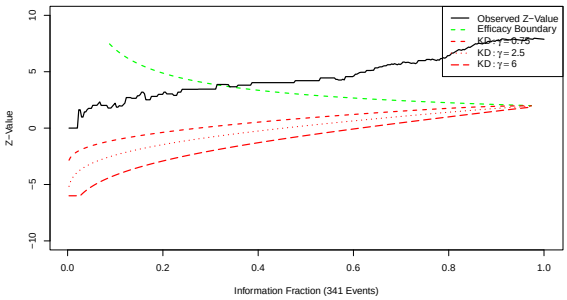
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

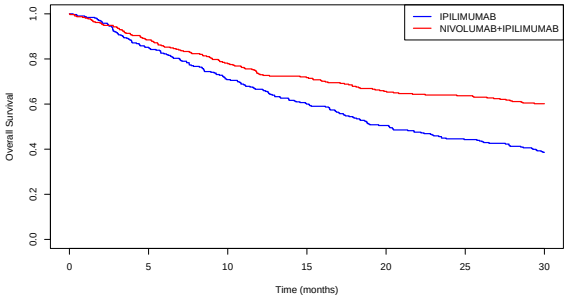


(g) *Piecewise Weibull Fit*

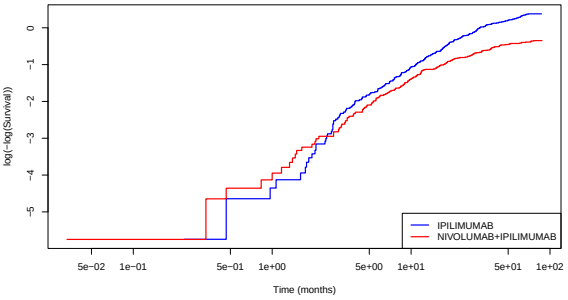


(h) *GSD Boundary Plot*

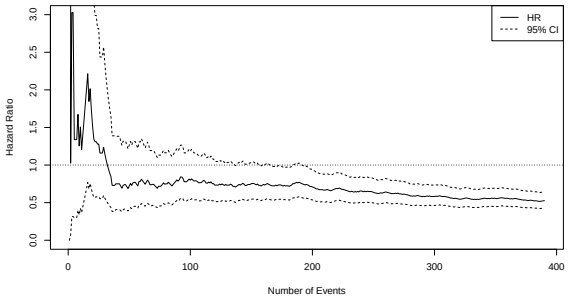
Trial D1 – Overall Survival (OS)



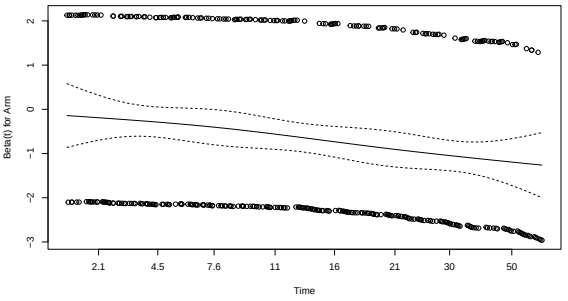
(a) *KM Plot*



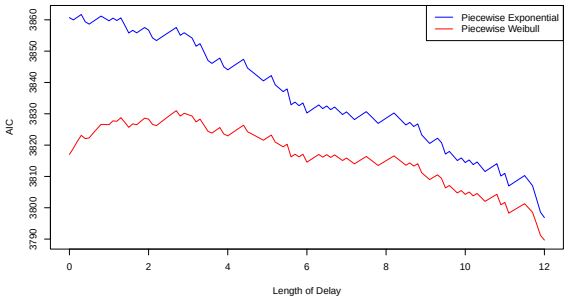
(b) *C-loglog Plot*



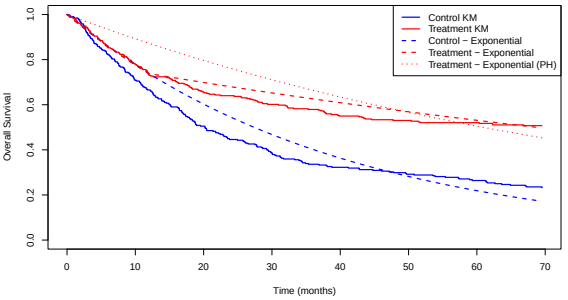
(c) *Hazard Ratio vs Events*



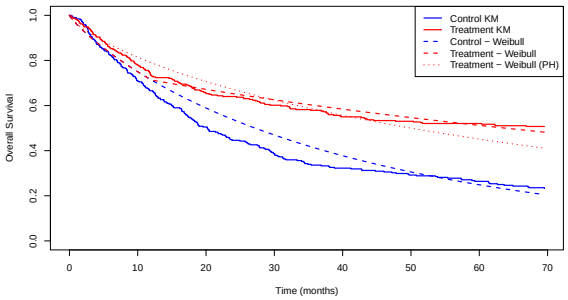
(d) *Schoenfeld residuals*



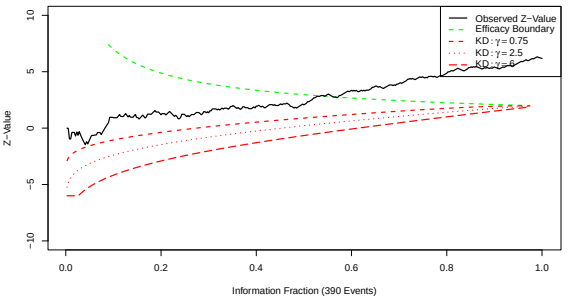
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

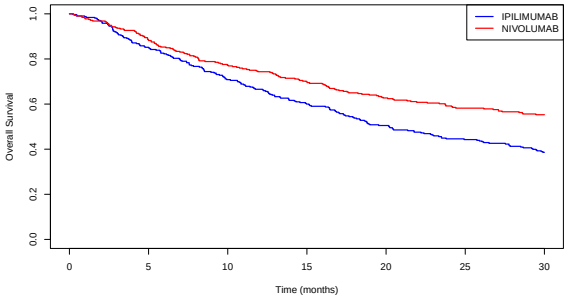


(g) *Piecewise Weibull Fit*

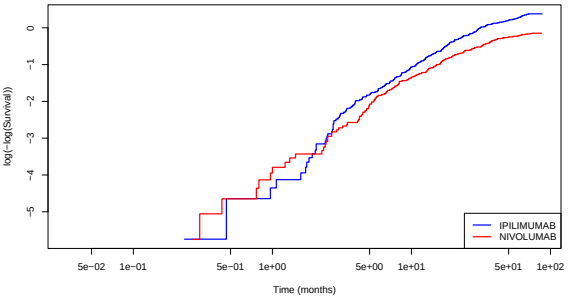


(h) *GSD Boundary Plot*

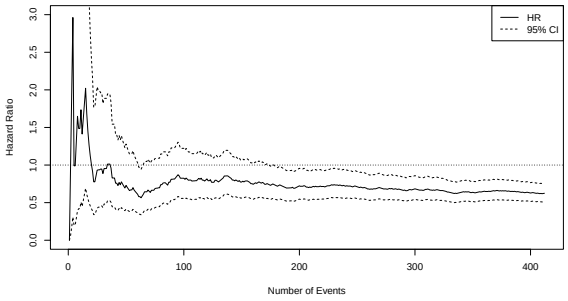
Trial D2 – Overall Survival (OS)



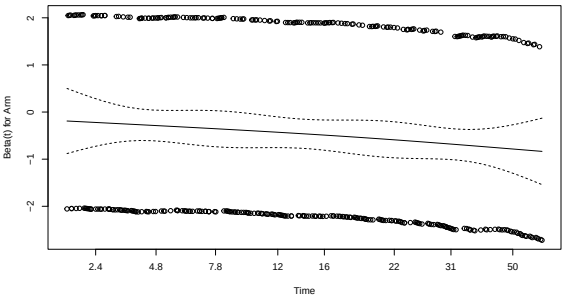
(a) *KM Plot*



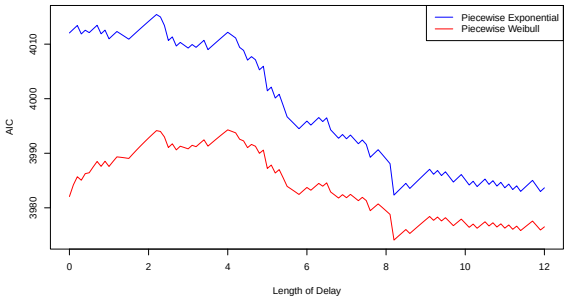
(b) *C-loglog Plot*



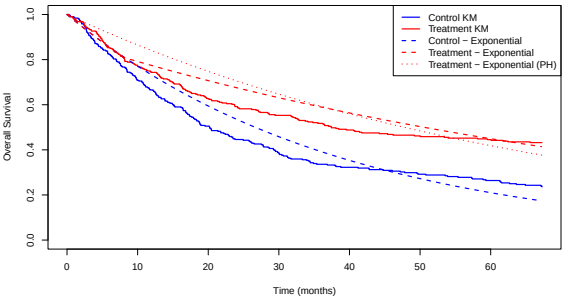
(c) *Hazard Ratio vs Events*



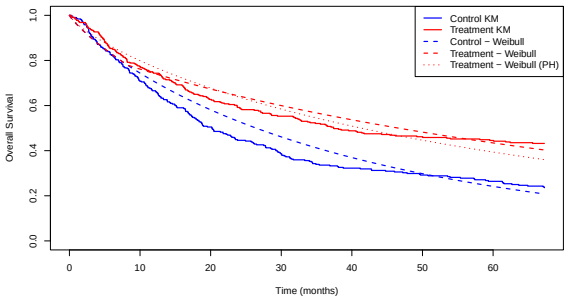
(d) *Schoenfeld residuals*



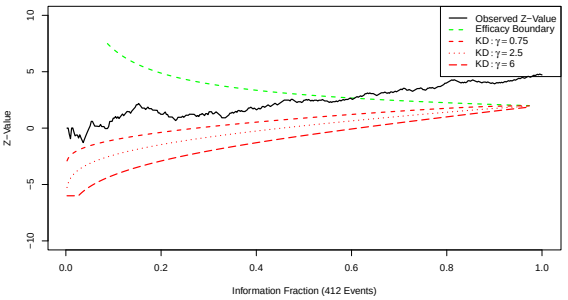
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

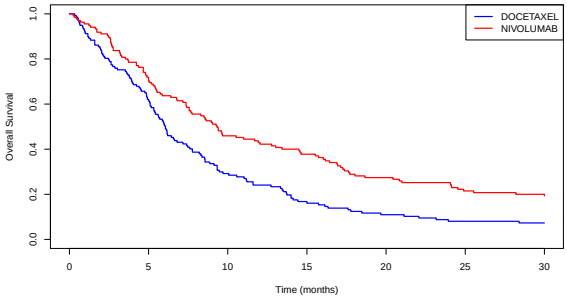


(g) *Piecewise Weibull Fit*

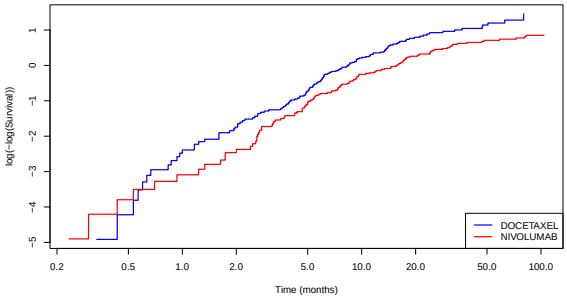


(h) *GSD Boundary Plot*

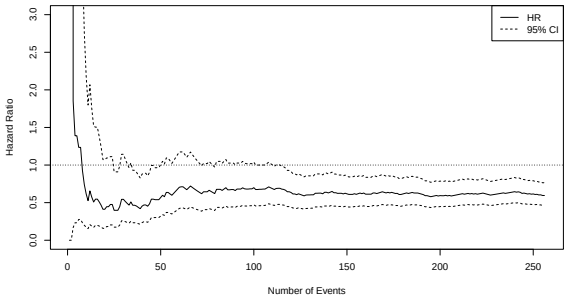
Trial E – Overall Survival (OS)



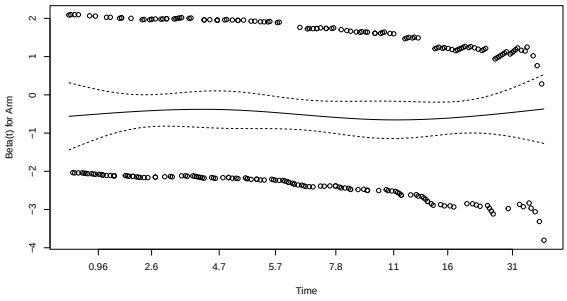
(a) *KM Plot*



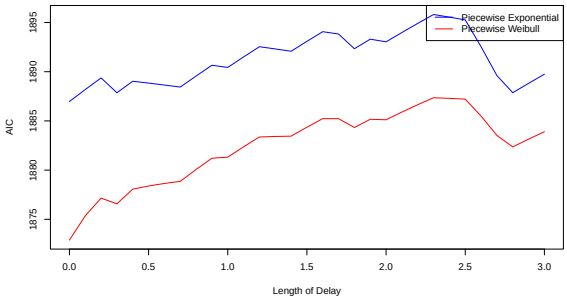
(b) *C-loglog Plot*



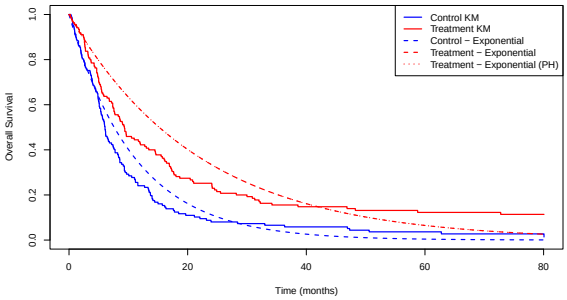
(c) *Hazard Ratio vs Events*



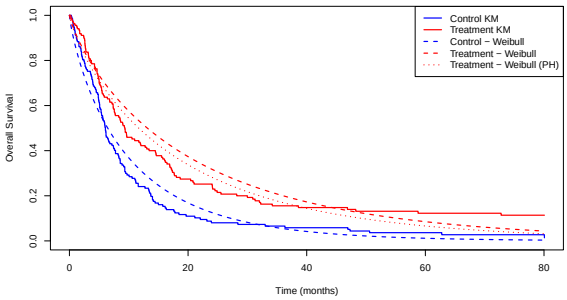
(d) *Schoenfeld residuals*



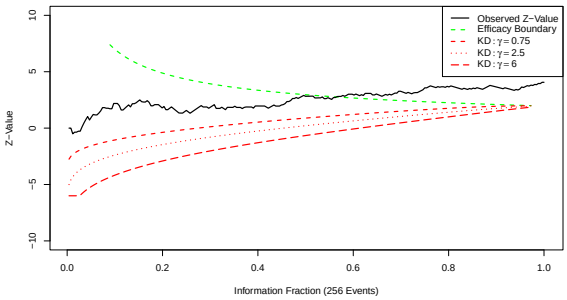
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

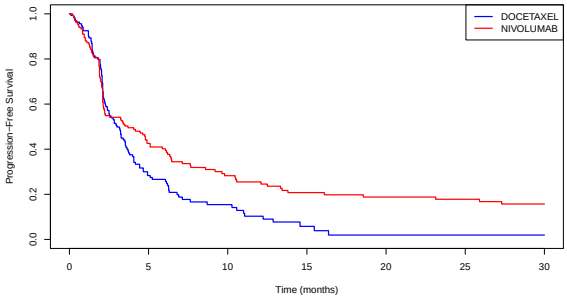


(g) *Piecewise Weibull Fit*

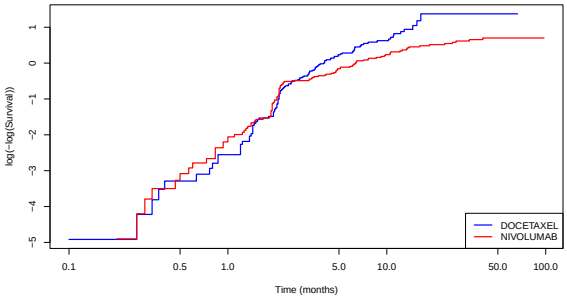


(h) *GSD Boundary Plot*

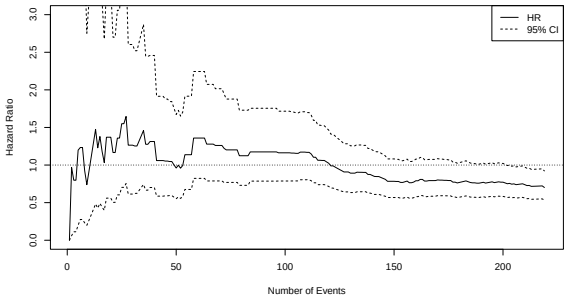
Trial E – Progression-Free Survival (PFS)



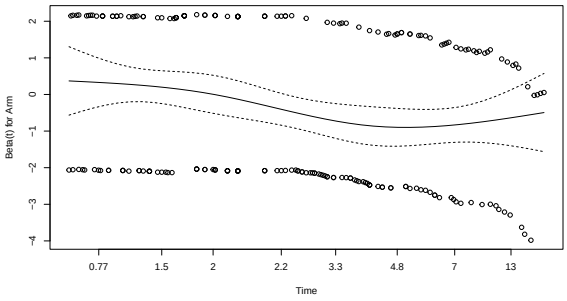
(a) *KM Plot*



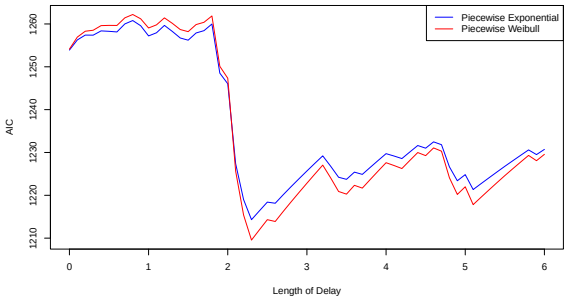
(b) *C-loglog Plot*



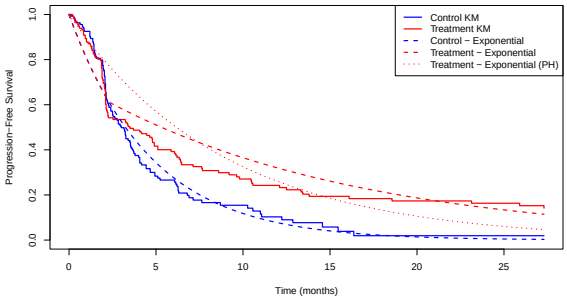
(c) *Hazard Ratio vs Events*



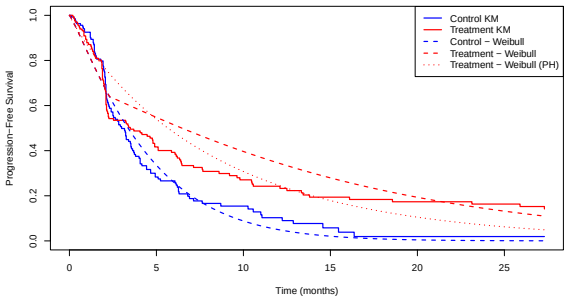
(d) *Schoenfeld residuals*



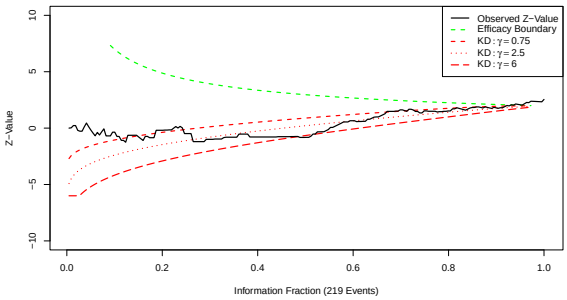
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

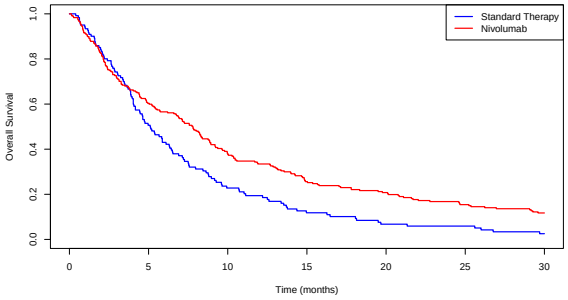


(g) *Piecewise Weibull Fit*

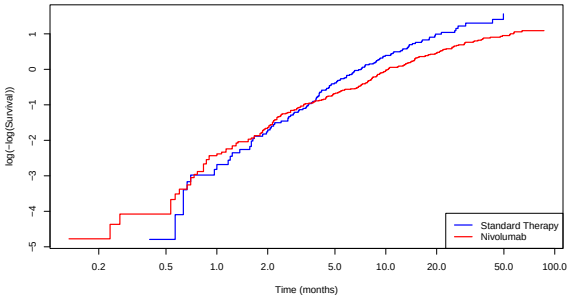


(h) *GSD Boundary Plot*

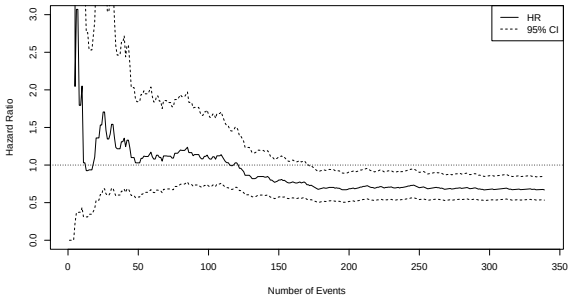
Trial G – Overall Survival (OS)



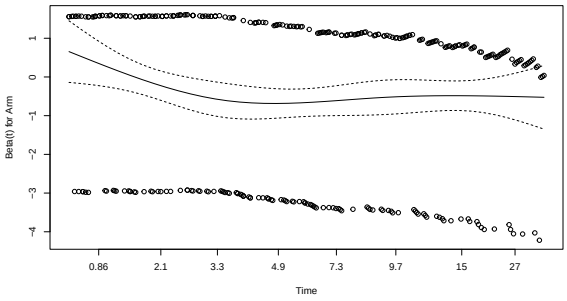
(a) *KM Plot*



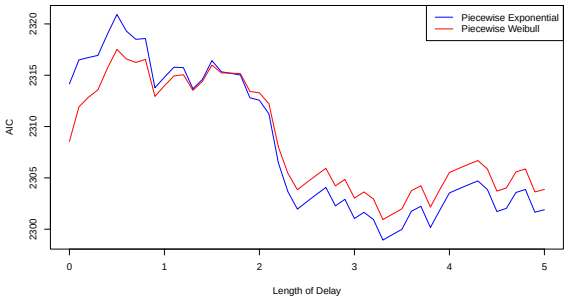
(b) *C-loglog Plot*



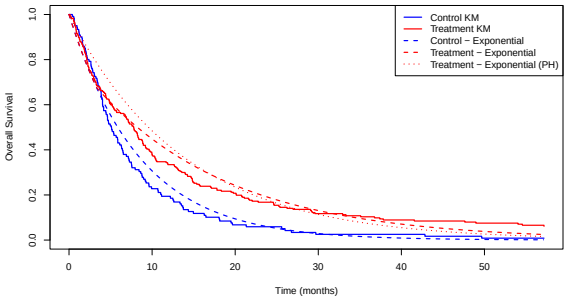
(c) *Hazard Ratio vs Events*



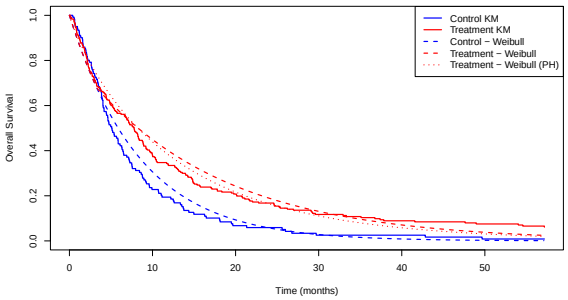
(d) *Schoenfeld residuals*



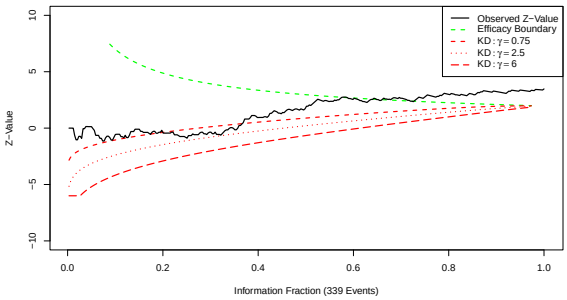
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

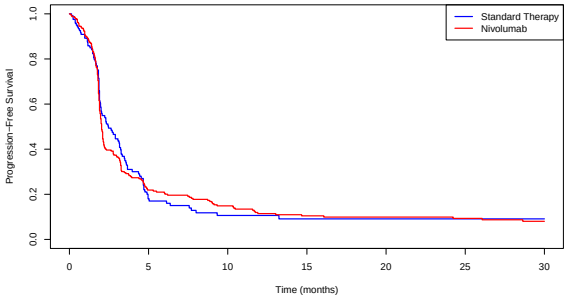


(g) *Piecewise Weibull Fit*

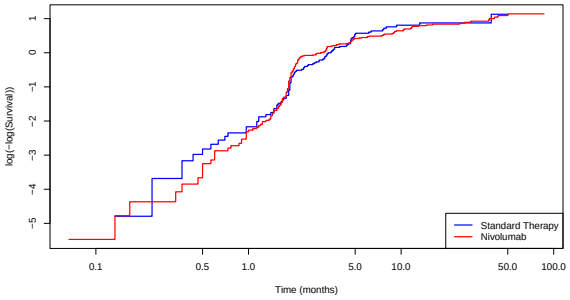


(h) *GSD Boundary Plot*

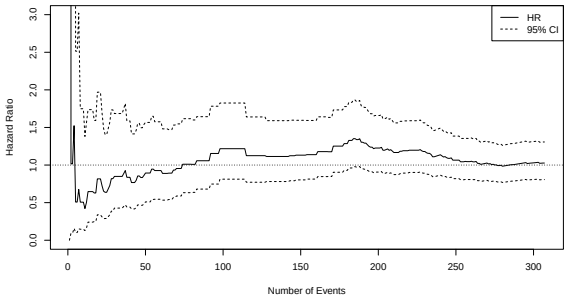
Trial G – Progression-Free Survival (PFS)



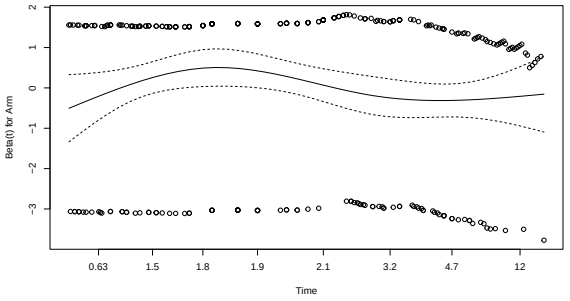
(a) *KM Plot*



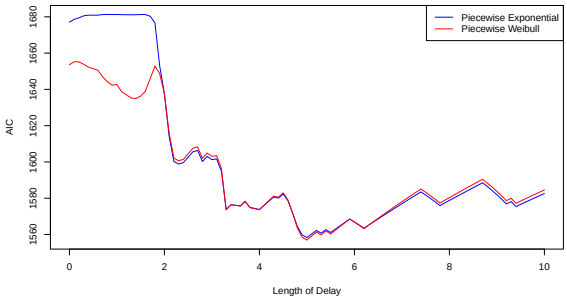
(b) *C-loglog Plot*



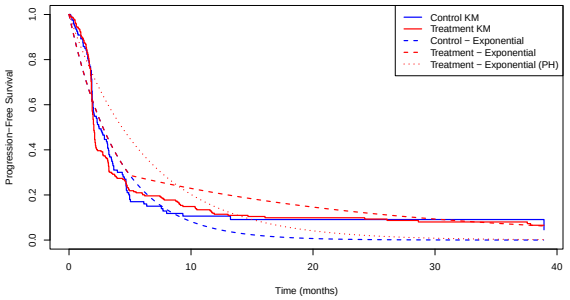
(c) *Hazard Ratio vs Events*



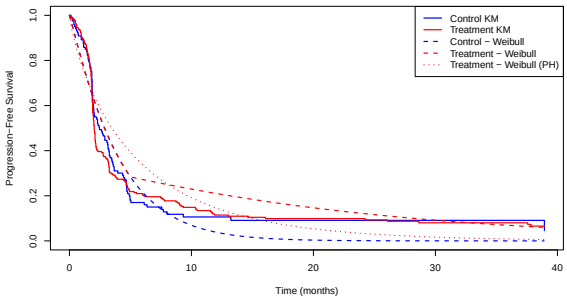
(d) *Schoenfeld residuals*



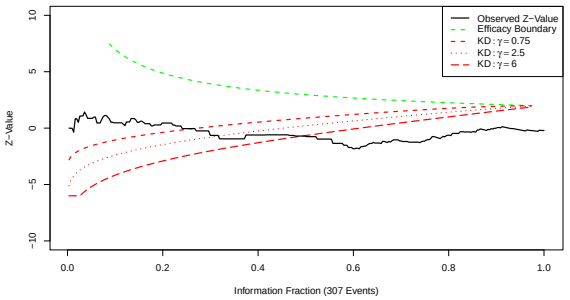
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

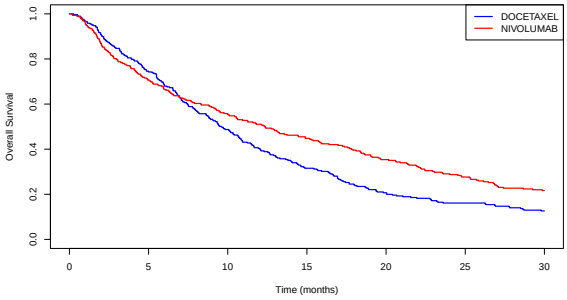


(g) *Piecewise Weibull Fit*

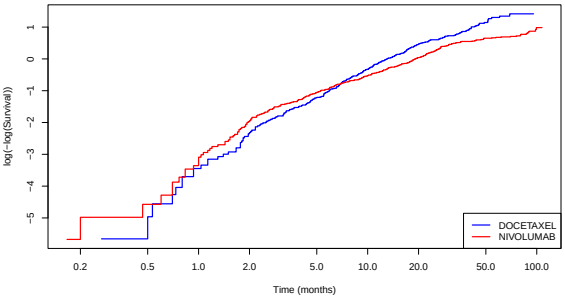


(h) *GSD Boundary Plot*

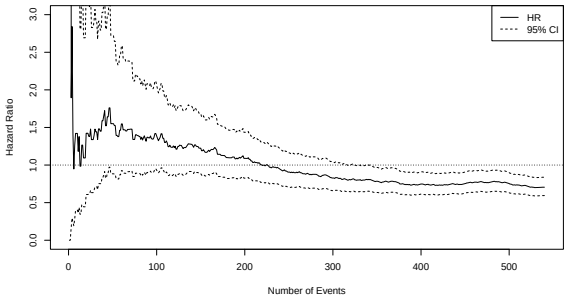
Trial H – Overall Survival (OS)



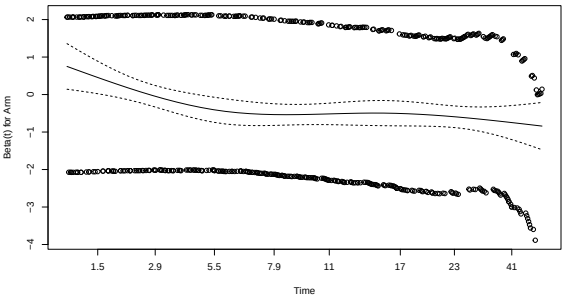
(a) *KM Plot*



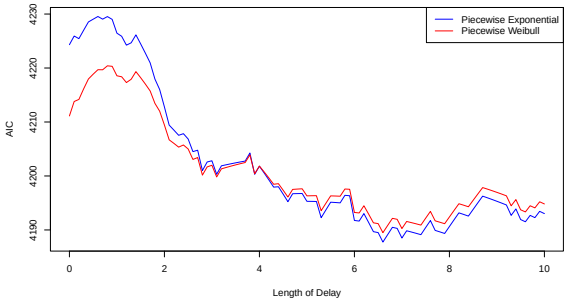
(b) *C-loglog Plot*



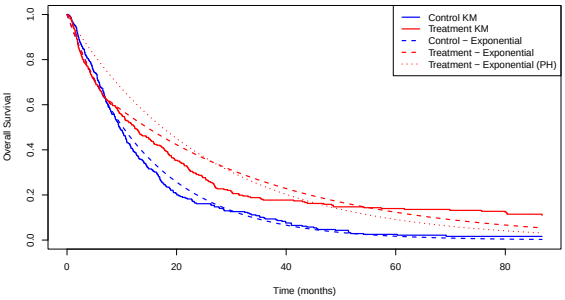
(c) *Hazard Ratio vs Events*



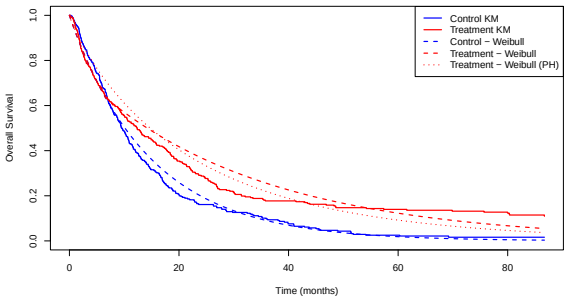
(d) *Schoenfeld residuals*



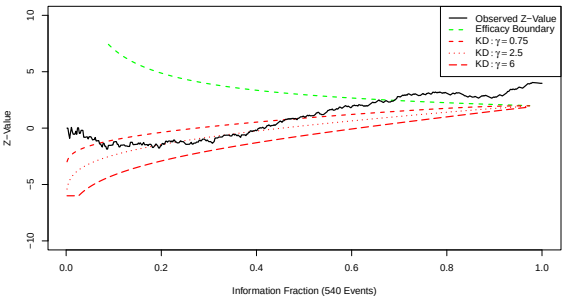
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*

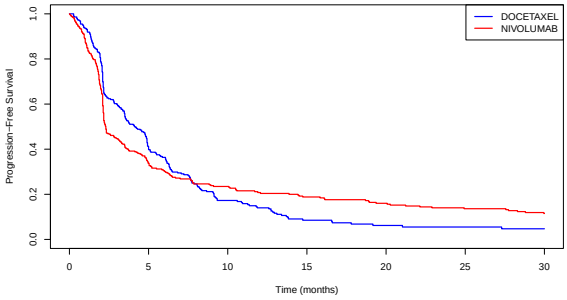


(g) *Piecewise Weibull Fit*

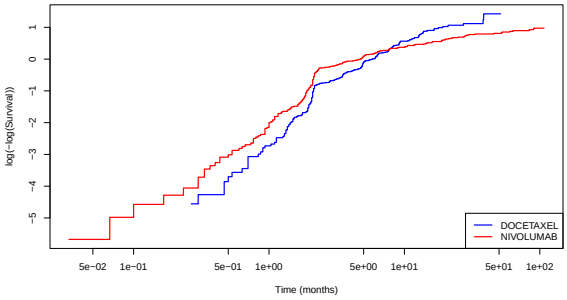


(h) *GSD Boundary Plot*

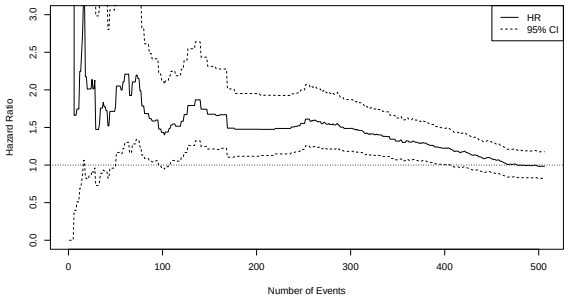
Trial H – Progression-Free Survival (PFS)



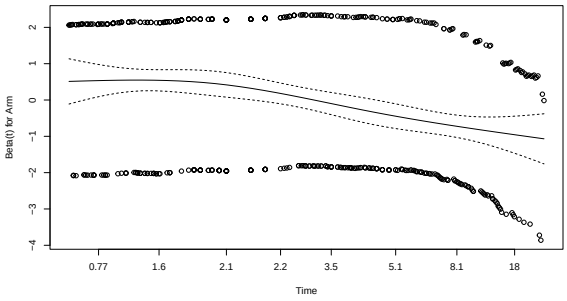
(a) *KM Plot*



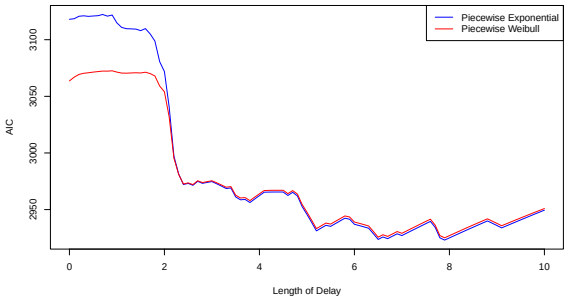
(b) *C-loglog Plot*



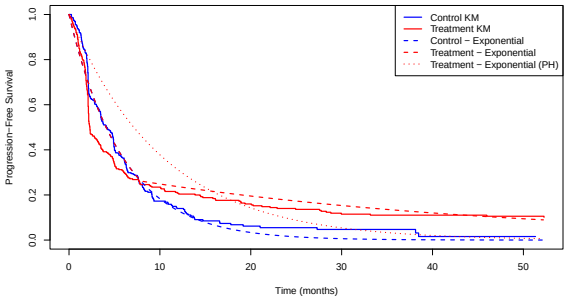
(c) *Hazard Ratio vs Events*



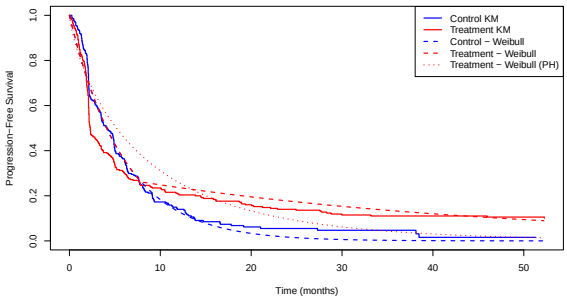
(d) *Schoenfeld residuals*



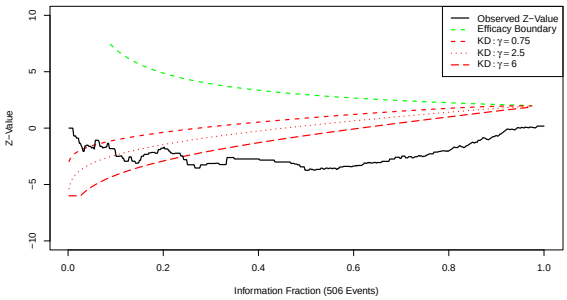
(e) *AIC vs Length of Delay*



(f) *Piecewise Exponential Fit*



(g) *Piecewise Weibull Fit*



(h) *GSD Boundary Plot*

Table C.3: *Interim stopping behaviour for PFS for the trials under one-sided O'Brien–Fleming ($\alpha = 0.025$) efficacy spending and Kim–DeMets ($\beta = 0.10$) futility spending. First crossing reported for each rule; futility treated as non-binding in this diagnostic analysis.*

Trial	Futility rule	Futility crossed?	Efficacy crossed?	Comment
B	$\gamma = 0.75$	132 (17.8%)	–	
	$\gamma = 2.5$	167 (22.4%)	–	
	$\gamma = 6$	290 (38.9%)	–	
C	$\gamma = 0.75$	–	107 (31.4%)	
	$\gamma = 2.5$	–	107 (31.4%)	
	$\gamma = 6$	–	107 (31.4%)	
E	$\gamma = 0.75$	25 (11.4%)	204 (93.2%)	
	$\gamma = 2.5$	58 (26.5%)	204 (93.2%)	
	$\gamma = 6$	107 (48.9%)	204 (93.2%)	
G	$\gamma = 0.75$	82 (26.7%)	–	
	$\gamma = 2.5$	98 (31.9%)	–	
	$\gamma = 6$	152 (49.5%)	–	
H	$\gamma = 0.75$	16 (3.2%)	–	
	$\gamma = 2.5$	51 (10.1%)	–	
	$\gamma = 6$	123 (24.3%)	–	