Some group sequential trials from industry over the last 30 years

Keaven M. Anderson [*]
Merck & Co., Inc., Rahway, NJ, USA
and
Yujie Zhao
Merck & Co., Inc., Rahway, NJ, USA
and
Nan Xiao
Merck & Co., Inc., Rahway, NJ, USA
and
Joy Ge
Merck & Co., Inc., Rahway, NJ, USA
and
Harlan F. Weisman
And-One Consulting, LLC

December 14, 2023

Abstract

We consider several industry group sequential trials and associated issues over the last 30 years. Generally, group sequential design has provided a great deal of flexibility to overcome many challenges in a relatively straightforward way compared to more complex adaptive designs. Among the issues considered are the timing of and boundaries for interim and final analyses, dealing with multiple hypotheses created by dose groups, populations and endpoints. Tools for design and execution will also be discussed.

Keywords: group sequential design, clinical trials, futility, adaptive design

[*] The authors gratefully acknowledge the many contributors to trials discussed in this paper as well as those contributing to the software packages used.

1 Introduction

Group sequential design enables analysis of accumulating data from a clinical trial as it is ongoing. There are many summaries of the use of such designs, e.g., Ellenberg et al. (2019). As noted there, group sequential trials became common at the National Institutes of Health in the 1970's, in cancer clinical trials in the 1980's and in industry trials in the 1990's. In this article, we review industry experience that we hope provides practical suggestions on the value, design and implementation of group sequential design. While the full sample size for such a trial should enable detection of minimally important clinical differences in both safety and efficacy outcomes, interim analyses can ensure earlier decision making when important larger emerging differences in safety or efficacy are observed. An independent Data Monitoring Committee (DMC) is often used to evaluate any such emerging differences. Among the responsibilities of a DMC evaluating interim analysis data as outlined by Ellenberg et al. (2019) the primary ones "*are to (i) safeguard the interests of study patients, (ii) to preserve the integrity and credibility of the trial in order that future patients may be treated optimally; and (iii) to ensure that definitive and reliable results be available in a timely way to the medical community.*"

We consider several potential advantages of group sequential analysis of data. First, early futility analysis can be used to rule out important safety issues and/or to require a positive enough trend to establish proof-of-concept for a new treatment (Ellenberg and Shaw, 2022). Efficacy interim analysis can more promptly result in approval of highly efficacious treatments to fulfill critical unmet medical needs (e.g., Gandhi et al., 2018; Powles et al., 2020). Multiple hypothesis testing, while not unique to group sequential design, can provide large scale data on definitive endpoints for testing multiple arms (doses) or testing multiple populations (e.g., biomarker positive and overall). Group sequential design is a straightforward and well-understood form of adaptive design (Center for Biologics Evaluation and Research and Center for Drug Evaluation and Research, 2019). We take the position that group sequential design can be a quite flexible and effective form of adaptive design. We also provide some cautions on other forms of adaptive designs due to logistical considerations and the difficulties of homogeneity assumptions required for many adaptive design methods to perform as expected.

The software used for this paper is primarily the **gsDesign2** R package (Anderson et al., 2022a). This package allows flexibility beyond its predecessor **gsDesign** (Anderson, 2020) which is also used here. The additional features of **gsDesign2** include enabling design with non-proportional hazards assumptions for time-to-event endpoints and the ability to design based on stratified populations for both binary and time-to-event endpoints. More flexibility in whether to utilize both upper and lower bounds at each analysis is also provided. Finally, many alternative testing methods are included in **gsDesign2**; this includes design with the weighted logrank approaches of Magirr and Burman (2019), Magirr (2021), and Fleming and Harrington (2011), as well as the MaxCombo approach to test with both the logrank test and one or more Fleming-Harrington weighted logrank tests (Roychoudhury et al., 2021).

The remainder of the paper is divided into sections using past clinical trial experience to reflect on advantages and challenges of group sequential design. Prior to this we provide an abbreviated literature review. We conclude the paper with a discussion.

2 Literature Review and Methods

We provide a brief literature review without attempting to be comprehensive. We cite both frequentist and Bayesian approaches to group sequential design. More comprehensive reviews can be found in, for example, Jennison and Turnbull (1999), Proschan et al. (2006), Wassmer and Brannath (2016), and Emerson et al. (2007). Some readers may wish to skip or use the this section for reference for the later sections summarizing trial experiences.

2.1 Asymptotic Distribution Theory

Jennison and Turnbull (1999) focus on the *canonical form* for group sequential design. That is, there are K analyses with

i) test statistics $Z_1 \ldots, Z_K$ that are multivariate normal,

ii) $\mathrm{E}(Z_k) = \theta\sqrt{\mathcal{I}_k}, k = 1, \ldots, K$, and

iii) $\mathrm{Cov}(Z_i, Z_j) = \sqrt{I_i/I_j}$, $1 \leq i \leq j \leq K$.

The **gsDesign2** package uses a more general assumption than *ii*, namely, $\mathrm{E}(Z_k) = \theta_k\sqrt{\mathcal{I}_k}, k = 1, \ldots, k$ in order to accommodate non-proportional hazards. Here, however, we limit ourselves to *ii*. The canonical form is useful quite broadly as noted by Scharfstein et al. (1997): *this limiting distribution arises naturally when one uses an efficient test statistic to test a single parameter in a semiparametric or parametric model.* The parameter θ for most of the examples presented here is the difference in the underlying failure rate for patients treated with an control regimen and an experimental regimen:

$$\theta = p_T - p_C. \tag{1}$$

We assume $X_{Ti}, i = 1, \ldots, n_{TK}$ are independent Bernoulli random variables with probability of failure p_T that $X_{Ti} = 1$ and $X_{Ti} = 0$, otherwise. Similarly we assume $X_{Ci}, i = 1, \ldots, n_{CK}$ are independent Bernoulli random variables with probability of failure p_C. Now we assume $n_{T1} < n_{T2} < \ldots < n_{TK}$, $n_{C1} < n_{C2} < \ldots < n_{CK}$ are sample sizes for the treatment and control groups at analyses $1, 2, \ldots, K$, respectively. We let

$$\hat{p}_{Tk} = \frac{\sum_{i=1}^{n_{Tk}} X_{Ti}}{n_{Tk}},\ \hat{p}_{Ck} = \frac{\sum_{i=1}^{n_{Ck}} X_{Ci}}{n_{Ck}},\ \text{and } \hat{\theta}_k = \hat{p}_{Ck} - \hat{p}_{Tk}. \tag{2}$$

The variance of $\hat{\theta}_k = \hat{p}_{Ck} - \hat{p}_{Tk}$ is

$$\text{Var}(\hat{\theta}_k) = \text{Var}(\hat{p}_{Ck}) + \text{Var}(\hat{p}_{Tk}) = \frac{p_C(1-p_C)}{n_{Ck}} + \frac{p_T(1-p_T)}{n_{Tk}} = \mathcal{I}_k^{-1}. \tag{3}$$

When testing, we use estimates of p_C, p_T to obtain

$$Z_k = \hat{\theta}\sqrt{\hat{\mathcal{I}}_{0k}}. \tag{4}$$

We have generally

$$\hat{\mathcal{I}}_k^{-1} = \widehat{\text{Var}}(\hat{\theta}_k) = \frac{\hat{p}_{Ck}(1-\hat{p}_{Ck})}{n_{Ck}} + \frac{\hat{p}_{Tk}(1-\hat{p}_{Tk})}{n_{Tk}}. \tag{5}$$

Under the null hypothesis that $p_C = p_T$ we let the overall event rate be denoted by

$$\hat{p}_k = \frac{\sum_{i=1}^{n_{Ck}} X_{Ci} + \sum_{i=1}^{n_{Tk}} X_{Ti}}{n_{Ci} + n_{Ti}} \tag{6}$$

and the null hypothesis estimates of statistical information and variance are

$$\hat{\mathcal{I}}_{0k}^{-1} = \widehat{\text{Var}}_0(\hat{\theta}_k) = \left(\frac{1}{n_{Ck}} + \frac{1}{n_{Tk}}\right)\hat{p}_k(1-\hat{p}_k). \tag{7}$$

We assume $\mathcal{I}_K$ is the planned statistical information for the final analysis and define the planned information fraction at analysis k as $t_k = \mathcal{I}_k/\mathcal{I}_K, k = 1,\ldots,K$. We also let $t_0 = 0$. We define B-values (Proschan et al. (2006)) for $k = 1,\ldots,K$ as

$$B_k = \sqrt{t_k}Z_k \tag{8}$$

which behaves like a Brownian motion with drift in that

- B1: $B_1,\ldots,B_K$ have a multivariate normal distribution,

- B2: $\mathrm{E}(B_k) = \theta t_k$,
- B3: $\mathrm{Cov}(B_i, B_j) = t_i$ for $1 \leq i \leq j \leq K$,
- B4: $B_j - B_i$ is independent of B_i with $\mathrm{E}(B_j - B_i) = \theta(t_j - t_i)$, $\mathrm{Var}(B_j - B_i) = (t_j - t_i)$ for $1 \leq i \leq j \leq K$.

B4 is the so-called independent increments property which we use below to simplify computation of conditional error and conditional power.

2.2 Rejection and Acceptance Regions

Bounds for group sequential designs can be characterized in various ways. For the most part, these can be considered monotone transformations of scale that have different interpretations. Here we will define bounds for group sequential testing on the B-value and Z-value scales. Suppose we set a lower and upper B-value bounds $-\infty \leq l_k < u_k \leq \infty, 1 \leq k < K$ and $-\infty \leq l_K \leq u_K < \infty$. On the Z-scale this translates $a_k = \sqrt{t_k} l_k, b_k = \sqrt{t_k} u_k, 1 \leq k \leq K$. We assume that at least one of l_k, u_k is finite for each $k \leq K$.

Figure 1 shows on the Z-scale how bounds may differ for different types of trials. We assume a 2-arm trial in all cases; we will refer to the arms as control and experimental here. 1) A 1-sided design ($a_k = -\infty, 1 \leq k \leq K$) has only an efficacy bound used to control Type I error for declaring experimental treatment superior to control assuming no difference. 2) A 2-sided symmetric design controls Type I error for declaring either control or experimental treatment superior assuming no difference. 3) A 2-sided asymmetric design with futility will be used here to mean a trial with a lower bound that controls Type II error during the course of the trial. A futility bound is generally intended to be used to stop a trial when efficacy at an interim analysis is insufficient to suggest a positive final finding. For the example shown, there is no futility analysis at the third interim due to

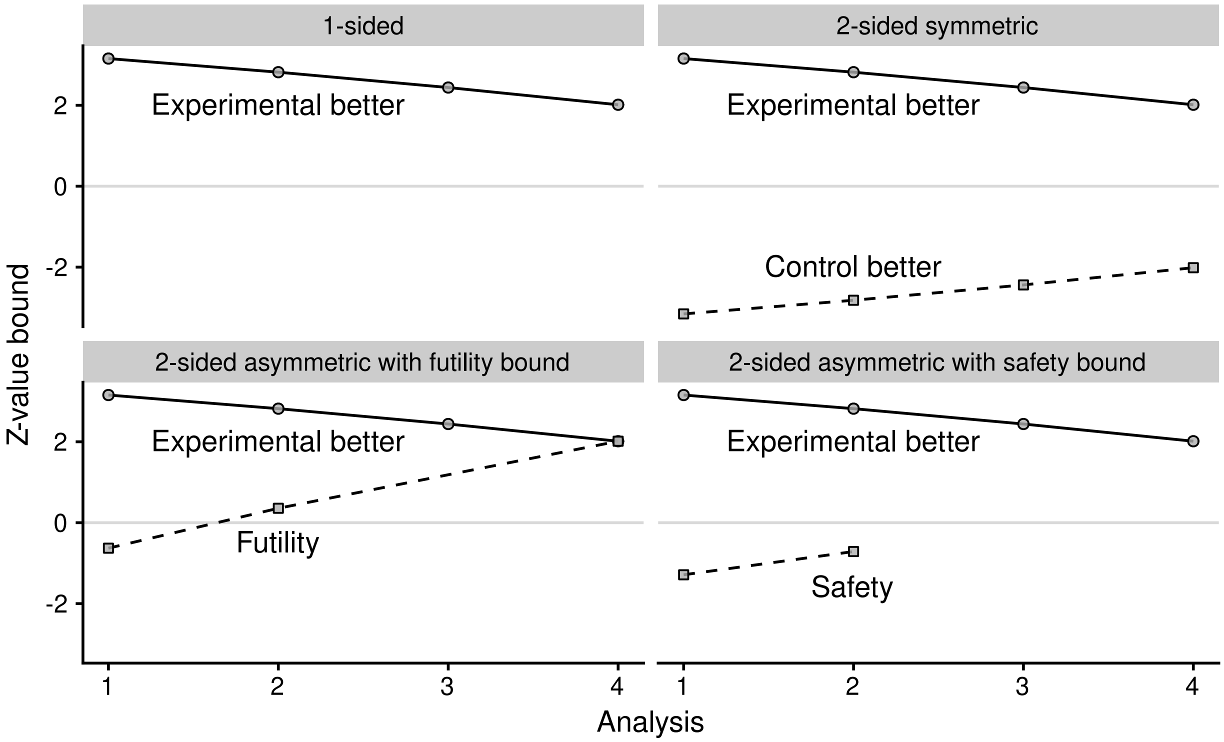

Figure 1: Different types of group sequential designs according to type of lower bound used.

the presumption that enrollment will have already proceeded to the final sample size prior to this interim analysis. We still show a futility bound at the final analysis which may control Type II error under the alternate hypothesis at a targeted level. 4) A 2-sided asymmetric with safety bound design will be used here to only stop for the lower bound if there is evidence to reject the null hypothesis in favor of control at an interim analysis. This futility bound is designed to control Type I error to stop in favor of control at a more relaxed level than stopping to declare superiority of experimental treatment. We exclude a safety evaluation at the third interim ($l_3 = -\infty$) for the same reason as the asymmetric 2-sided design. We exclude a safety bound at the final analysis ($l_4 = -\infty$) as control of Type I error of the trial overall is not an objective at that point. Safety bounds may be particularly useful when there is some suggestion that there may be a delayed effect onset or a subgroup that has high early risk of events in the experimental group.

We define the rejection region for testing the null hypothesis $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ at analysis $k = 1, \ldots, K$

$$R_k = \{B_k \geq u_k\} \cap_{j=1}^{k-1} \{l_j \leq B_j < u_j\}. \tag{9}$$

The region $A_k, 1 \leq k \leq K$ will be termed the acceptance region for now, but we have discussion of different uses for this region above and will further elaborate below.

$$A_k = \{B_k < l_k\} \cap_{j=1}^{k-1} \{l_j \leq B_j < u_j\}. \tag{10}$$

2.3 Type I Error and Power

Assuming the trial is stopped if a rejection or acceptance region is reached, we compute the probability of first rejecting the null hypothesis at analysis $k = 1, \ldots, K$ as

$$\alpha_k(\theta) = P_\theta(R_k). \tag{11}$$

The probability of first accepting the null hypothesis (futility bound), rejecting the null in favor of control (2-sided symmetric lower bound) or accepting there is a safety issue (safety bound) at analysis $k = 1, \ldots, K$ given a treatment effect θ is

$$\beta_k(\theta) = P_\theta(A_k). \tag{12}$$

We define further the non-binding rejection region preferred by regulators for group sequential design (Center for Biologics Evaluation and Research and Center for Drug Evaluation and Research (2019)) ignores any lower bound for Type I error computation:

$$R_k^+ = \{B_k \geq u_k\} \cap_{j=1}^{k-1} \{B_j < u_j\}. \tag{13}$$

Non-binding Type I error analysis $k = 1, \ldots, K$ is defined as

$$\alpha_{0k}^+ = P_{\theta=0}(R_k^+). \tag{14}$$

Total non-binding Type I error for a set of group sequential tests will be denoted as

$$\alpha_0^+ = \sum_{k=1}^{K} \alpha_{0k}^+. \tag{15}$$

Power, on the other hand, will be defined accounting for the lower Type I error for the sequence of tests as:

$$\alpha(\theta) = \sum_{k=1}^{K} P_\theta(R_k) = \sum_{k=1}^{K} \alpha_k(\theta). \tag{16}$$

For asymmetric designs with $l_K = u_K$, the Type II error for a given $\theta > 0$ under the restriction that $l_k = u_k$ is

$$\beta(\theta) = \sum_{k=1}^{K} P_\theta(A_k) = \sum_{k=1}^{K} \beta_k(\theta). \tag{17}$$

When $l_K < u_k$, $\beta(\theta)$ this is just the probability that a lower bound is crossed prior to crossing an efficacy bound (e.g., 2-sided symmetric design or asymmetric design with safety bound). For symmetric 2-sided trials with $l_k = -u_k, 1 \leq k \leq K$, $\beta(0)$ is the probability of rejecting the null hypothesis in favor of control benefit when there is no underlying treatment difference. Computations for Type I error, power and Type II error are all based on multiple integration; Jennison and Turnbull (1999), Chapter 19 provides these calculations which have been implemented in the **gsDesign** R package and generalized for non-proportional hazards in the **gsDesign2** R package. These computations include computation of boundary crossing probabilities as well as deriving bounds that have targeted boundary crossing probabilities. We will use the convention in the rest of this paper that if an asymmetric lower bound is derived to control boundary crossing probabilities under the null hypothesis $(\beta_k(0), k = 1, 2, \ldots, K)$ we will refer to the lower bound as a safety bound as we are trying to rule out worse outcomes in the experimental group than in control. Alternatively, if we are setting lower boundaries based Type II error under the alternative hypothesis $\theta = \theta_1 > 0$ $(\beta_k(\theta_1))$, we will refer to the bound as a futility bound as crossing such a bound has low probability under the meaningful treatment effect θ_1. For 2-sided symmetric designs, 2-sided Type I error for analysis k is computed as $\alpha_k(0) + \beta_k(0) = 2 \times \alpha_k(0), 1 \leq k \leq K$.

2.4 Bound Computation

There are several ways to compute bounds $l_k, u_k, k = 1, \ldots, K$ that can control Type I error at a targeted level. We provide an abbreviated summary here compared to Emerson et al. (2007).

- Haybittle (1971) and Peto et al. (1976) proposed symmetric, 2-sided interim Z-bounds $a_k = -b_k = 3$ which yields a 2-sided probability of crossing at a given interim analysis at $0.0027 = 2 \times 0.00135$. With for equally-spaced analyses and a final bound at a nominal 2-sided test at the 0.05 level, total Type I error is 0.0525, higher than the targeted 0.05. Changing the final bound to a 2-sided p=0.0474 bound controls Type I error at the targeted 0.05, 2-sided. This latter adjustment can be generalized and is referred to as a modified Haybittle-Peto bound.
- Slud and Wei (1982) also proposed 2-sided symmetric bounds, but chose arbitrary $\alpha_k(0) = \beta_k(0), k = 1, \ldots, K$ such that 2-sided $\alpha = \sum_{k=1}^{K}(\alpha_k(0) + \beta_k(0))$. It is straightforward using standard numerical integration tools to generate asymmetric bounds not requiring $\alpha_k(0) = \beta_k(0), k = 1, \ldots, K$.
- Wang and Tsiatis (1987) proposed boundary families where $l_k = u_k$ and

$$\sqrt{t_k}u_k = \Gamma(\alpha, K, \Delta)k^{\Delta} \tag{18}$$

where $\Gamma(\alpha, K, \Delta)$ is an appropriately defined constant to control Type I error. These bounds included both the classic aggressive Pocock (1977) bounds ($\Delta = 0.5$; $u_1\sqrt{t_1} =$

$u_2\sqrt{t_2} = \ldots = u_K\sqrt{t_K}$; note $t_K = 1$) and conservative O'Brien and Fleming (1979) bounds ($\Delta = 0$ with u_k constant in k; on the Z-statistic scale, bounds are decreasing in k). While this was proposed for symmetric bounds with equally-spaced analyses, 1-sided and/or unequally-spaced analyses can be derived with this approach.

- Pampallona and Tsiatis (1994) generalized the Wang-Tsiatis bounds by setting upper bounds as above to control one-sided Type I error, but setting futility bounds to control lower boundary crossing probabilities for some $\theta_1 > 0$ at level $0 < \beta < 1 - \alpha$ using

$$\sqrt{t_k}l_k = k\theta_1 - \Gamma_2(\alpha, \beta, K, \Delta)k^{\Delta}. \qquad (19)$$

- Lan and DeMets (1983) proposed bounds based on specifying $\alpha_k(0) = \beta_k(0) = \alpha^*(t_k)$ for an increasing function $\alpha^*(t), t \geq 0$ with $\alpha^*(0) = 0$, and for $t \geq 1, \alpha^*(t) = \alpha$, the 2-sided Type I error targeted. This allows a wide variety of boundary types. These can also be used as 1-sided efficacy bounds with no futility stopping. Lan and DeMets (1983) propose the spending functions to approximate O'Brien and Fleming (1979) bounds where for $0 < t \leq 1$

$$\alpha^*(t_k) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t}) \qquad (20)$$

where $\Phi()$ is the standard normal cumulative distribution function as well as Pocock (1977) bounds where

$$\alpha^*(t_k) = \min(\alpha \log[1 + (e - 1)t], \alpha). \qquad (21)$$

The O'Brien-Fleming-like spending function in (20) is often preferred by regulators to require stringent efficacy bounds for early stopping in Phase III trials planned for approval of a new treatment. As noted below in our discussion of Bayesian boundaries, bounds based on Pocock-like spending can be useful for early-stage trials to make prompt development decisions when there is little prior information on treatment effect. Anderson and Clark (2010) summarize a variety of spending functions and suggest fit-for-purpose spending as will be described in an example below. These include the power spending function of Kim and DeMets (1987), another commonly used flexible spending family by Hwang et al. (1990); more recently Xi and Gallo (2019) suggested a spending approach related to conditional error.

- Pampallona et al. (2001) introduced β-spending functions $\beta^*(t)$ that increase from 0 when $t = 0$ to β (Type II error) or $1 -$ power under some alternative $\theta = \theta_1 > 0$. These futility bounds are generally set up so that $l_K = u_K$ and for $1 \leq k \leq K, -\infty < l_k < u_k < \infty$. In this case, for $k = 1, \ldots, K$ we let $\beta_k(\theta_1) = \beta^*(t_k) - \beta^*(t_{k-1})$ to back-calculate bounds that have controlled Type II error increments at each analysis. Choice of sample size is made to ensure $l_K = u_K$ and $\beta(\theta)$ is the Type II error.

- Asymmetric 2-sided bounds based on $\beta_k(0), k = 1, \ldots, K$ can be computed. This could emphasize early stopping if there is a trend in favor of control that is not as strong as might be required to stop early for an experimental treatment benefit. We will refer to such a bound here as a safety bound.

- While many of the above methods were originally developed for 2-sided testing, our experience with industry trials have been that the testing is one-sided by nature. Decision-making is asymmetric depending on whether a control or experimental treatment is favored. Thus, we tend to think in terms of one-sided testing. It is good to specify 1- vs. 2-sided testing up front.

- Finally, there are wedge bounds for early stopping for equivalence or failure to stop for equivalence; see, e.g., Jennison and Turnbull (1999) or Emerson et al. (2007) for further information as we will not discuss this approach further here.

In general, bounds based on spending functions are convenient as they adapt bounds to control operating characteristics when statistical information is not as planned at the time of analyses.

2.5 Conditional Error and Conditional Power

Conditional power and conditional error are defined by conditioning on a fixed θ-value and fixed interim test statistic:

$$
\begin{aligned}
CP_k(b,\theta) &= P_\theta\left(\bigcup_{j=k+1}^{K}\{R_j\}\;\middle|\;B_k=b\right)\\
&= P_\theta\left(\bigcup_{j=k+1}^{K}\{B_j\geq u_j\}\cap_{i=1}^{j-1}\{l_i\leq B_i<u_i\}\;\middle|\;B_k=b\right) \qquad (22)\\
&= P_\theta\left(\bigcup_{j=k+1}^{K}\{B_j-B_k\geq u_j-b\}\cap_{i=k}^{j-1}\{l_i-b\leq B_i-B_k<u_i-b\}\right).
\end{aligned}
$$

The transition from line 2 to line 3 of (22) is based on the memoryless property B4 of section 2. While conditional power power has typically been computed at $\theta=\hat{\theta}$, Liu and Chi (2001) suggested using $\theta=\theta_1$, the θ-value for which a design is powered. Conditional error based on non-binding futility can be defined as

$$
\begin{aligned}
CP_k^{+}(b,0) &= P_0\left(\bigcup_{j=k+1}^{K}\{R_j^{+}\}\;\middle|\;B_k=b\right)\\
&= P_0\left(\bigcup_{j=k+1}^{K}\{B_j-B_k\geq u_j-b\}\cap_{i=k}^{j-1}\{B_i-B_k<u_i-b\}\right).
\end{aligned} \qquad (23)
$$

These forms of conditional power and conditional error use the same numerical integration forms as for power, Type I error and Type II error computations above. Conditional error is particularly useful for adaptive design as has been noted by Müller and Schäfer (2001); this will not be further discussed here.

2.6 Bayesian Analyses

Two approaches to Bayesian analysis have been suggested. In either case, a prior distribution for a single parameter such as θ or individual distribution parameters such as p_T, p_C above could be considered. For the latter, we refer the reader to Gerber and Gsponer (2016) who provide useful examples and software (the **gsbDesign** R package). Here, we consider a prior distribution for θ and will stick with the asymptotic formulation (canonical form) for tests described above. While the **gsDesign** package enables arbitrary prior distributions, we consider a conjugate normal prior as suggested in, for example, Freedman and Spiegelhalter (1989) in their comparison of frequentist and Bayesian bounds. We assume a prior distribution for θ:

$$\theta \sim \text{Normal}\left(\mu_0, \sigma^2)\right). \tag{24}$$

With an observed efficient estimate $\hat{\theta}_k$ at analysis k with statistical information $\mathcal{I}_k$, the posterior distribution for θ is

$$\theta \sim \text{Normal}\left(\frac{\mu_0/\sigma^2 + \hat{\theta}_k \mathcal{I}_k}{1/\sigma^2 + \mathcal{I}_k}, \left(1/\sigma^2 + I_k\right)^{-1}\right). \tag{25}$$

That is, the posterior distribution weights the prior and observed means by their respective statistical information (inverse variance). Freedman and Spiegelhalter (1989) use the assumption that for some σ_0, n_0, $\sigma^2 = \sigma_0^2/n_0$ and for $k = 1, \ldots, K, \mathcal{I}_k^{-1} = \sigma_0^2/n_k$. The posterior distribution in (25) becomes

$$\theta \sim \text{Normal}\left(\frac{\mu_0 n_0 + \hat{\theta}_k n_k}{n_0 + n_k}, \frac{\sigma_0^2}{n_0 + n_k}\right). \tag{26}$$

We now set posterior bounds for efficacy at

$$P_{\text{posterior}}(\theta > \theta_E) = \epsilon_e \tag{27}$$

and to stop for futility if

$$P_{\text{posterior}}(\theta < \theta_f) = \epsilon_f. \tag{28}$$

Freedman and Spiegelhalter (1989) note further that for a weak prior (n_0 small and $\theta_0 = 0$), these bounds are comparable to the aggressive early stopping Pocock (1977) bounds, while for a strong prior (n_0 large and $\theta_0 = 0$), the bounds are comparable to O'Brien and Fleming (1979) which are conservative for early stopping.

Another approach to analysis is to compute posterior predictive power. This measure averages conditional power over the posterior distribution for treatment effect given an interim result. Evaluating this for observed outcomes can be a useful alternative to conditional power. This can also be used to describe bounds $l_k, u_k, k < K$. The posterior predictive power based on an observed test statistic Z_k and corresponding interim treatment effect estimate $\hat{\theta}_k$ is computed as

$$p_{\text{predictive}}(Z_k, \hat{\theta}_k) = \int_{-\infty}^{\infty} f_{\text{posterior}}(\theta \mid \hat{\theta}_k) CP(Z_k \mid \theta) d\theta. \tag{29}$$

3 The EPIC trial

The EPIC Investigators (1994) studied participants undergoing percutaneous transluminal coronary artery angioplasty (PTCA) at high risk for recurrent events. This double-blind study was designed to determine the safety and efficacy of abciximab or placebo in this indication when added to standard therapy with aspirin and heparin. The adjudicated composite primary endpoint was a binary outcome that included the components 1) death, 2) nonfatal myocardial infarction, and 3) urgent repeat intervention (PTCA or coronary bypass surgery). Potential safety concerns at the time of study design were major bleeding, including a relatively rare but severe event, intracranial hemorrhage. The EPIC design included 3 treatment arms:

- Abciximab bolus and infusion: Abciximab intravenous bolus and intravenous infusion
- Abciximab bolus: Abciximab intravenous bolus and intravenous placebo infusion
- Placebo: placebo intravenous bolus and placebo intravenous infusion

The trial employed a group sequential design with planned enrollment of 2100 participants. This was powered to detect a reduction from 15% to 10% in the primary endpoint with 80% power and 2-sided $\alpha = 0.05$. Interim analyses were planned after 1/3 and 2/3 of participants were enrolled.

The final analysis bound required a nominal 2-sided $p \leq 0.036$. While the actual spending function was not published, this could be achieved using a Hwang-Shih-DeCani spending function with $\gamma = -4.9$. Multiplicity control for comparison of two experimental arms versus a common control was achieved by using a global null hypothesis trend test where the control was coded as 0, abciximab bolus group as 1, and the abciximab bolus and infusion as 2. The Mantel-Haenszel trend test (Mantel (1963)) used for this purpose gives little statistical leverage to the bolus only group and is, thus, similar to a pairwise comparison of abciximab bolus and infusion versus placebo. In any case, after a positive trend test, the pairwise comparison of each abciximab group versus control could be tested at the same α-level. This is a simple example of multiplicity control when testing multiple hypotheses; for a review of some related literature, see Anderson et al. (2022b).

Adjudication was considered necessary for accurate assessment of the primary endpoint in this trial. Prior to sending data to the central adjudication group, substantial data collection, cleaning and transfer were required. Set up of adjudication logistics and personnel can be a challenge and may be too slow for prompt availability for Data Monitoring Committee review at the time of interim analyses. On the other hand, the key component that eventually demonstrated the underlying primary endpoint benefit was the diagnosis of myocardial infarction; central adjudication for this endpoint component was particularly important. For the EPIC trial, the balance of safety and efficacy during the course of the trial was important. This is indicated by the eventual final summary of major bleeding versus efficacy results as summarized in the following table. These results were carefully examined and balanced versus efficacy at the time of interim analyses by the DMC to confirm the rationale to continue the trial. The final results yielded a positive efficacy finding with discouraging safety that may have limited the use of this innovative treatment. The *p*-values of 0.009 for the trend test, 0.008 for bolus and infusion vs control (35% reduction) yielded a positive efficacy finding. The third arm (bolus) which had been strongly suggested by regulators before the trial start was essential to find the minimally effective dose in the context of safety issues observed. Without this third arm, the question of whether a bolus only approach could have been safer and equally efficacious would not have been answered, possibly resulting in no regulatory approval until this question was studied.

EPIC lessons learned included 1) the importance of logistics and execution for interim analyses, 2) interim analyses are important for both efficacy and safety, and 3) more than 1 experimental arm can be essential in Phase 3 to find an appropriate balance between efficacy and safety.

Table 1: EPIC results at final analysis.

Endpoint	Placebo	Abciximab bolus	Abciximab bolus and infusion
N	696	695	708
Primary efficacy	89 (12.4%)	79 (11.4%)	59 (8.3%)
Major bleeding	46 (6.6%)	76 (10.9%)	99 (13.9%)

			Abciximab bolus and
Endpoint	Placebo	Abciximab bolus	infusion
Intracranial hemorrhage	2 (0.3%)	1 (0.1%)	3 (0.4%)

4 EPILOG

The positive results of the EPIC trial enabled development of a safer and more efficacious dosing strategy as well as evaluating a broader, lower-risk population by the EPILOG Investigators (1997). The dose groups studied were 1) placebo and standard-dose weight-adjusted heparin; 2) abciximab and standard-dose weight-adjusted heparin; or 3) abciximab and low-dose weight-adjusted heparin. The primary endpoint of the trial was the same 30-day endpoint as in the EPIC trial. Type I error was controlled for testing of this endpoint and a 6-month composite endpoint of death, myocardial infarction, coronary bypass surgery or repeated percutaneous revascularization (urgent or nonurgent). For each of these endpoints, multiplicity was controlled by first testing for differences in the combined abciximab treatment groups versus placebo. If this test were positive, testing of individual abciximab treatment groups versus the placebo treatment group would proceed. Each of these tests was controlled by group sequential testing, anticipating the later multiplicity control work of Maurer and Bretz (2013). At the time of publication, the EPILOG Investigators (1997) used simulation to evaluate Type I error control.

The EPILOG trial provides a good example of sample size adaptation with group sequential design as well as a specific safety bound for lack of efficacy and the importance of study logistics. Enrollment began on February 27, 1995 and the trial was terminated on December 14, 1995 after an interim analysis of the first 1500 participants. The study revealed both a lower control event rate and a more substantial reduction in the primary endpoint (30-day endpoint rates: 8.2% in the control group compared to 3.6% and 2.6% in the two experimental arms).

By the time the trial was stopped following the interim analysis, a total of 2792 participants were enrolled. It is not uncommon that a protocol can have this type of accelerated enrollment and substantial over-enrollment beyond participants with cleaned data for an interim efficacy analysis. Understanding enrollment rates, follow-up requirements, data collection, data cleaning, adjudication and their relation to being able to review data in a timely fashion can be critical. If there is substantial over-enrollment between the data cutoff and analysis, any summary of later available data not included in the formal analysis should be discussed with the DMC prior to the unblinded review. Not noted in the publication was a study amendment based on blinded data review to substantially increase the final sample size to 4800. The blinded review of study endpoints revealed a substantially lower rate than in the previous EPIC trial that was in a higher-risk population. The group sequential design included an unbalanced bound at the first interim analysis of where a positive efficacy finding required a p-value ≤ 0.0005 if an abciximab group had a lower event rate than control, while a p-value of ≤ 0.025 in favor of control (safety bound) would stop the trial for an unfavorable finding.

The adaptations in EPILOG were an effective combination of information-based sample size adaptation based on blinded review of event rates, group sequential stopping rules for efficacy and multiple hypothesis testing. Blinded sample size adaptation can be substantially simpler and will generally receive less regulatory scrutiny than unblinded efficacy adaptation. An unblinded efficacy adaptation (e.g., Chen et al., 2004) would likely have resulted in the same positive efficacy finding without the substantial sample size increase generated by the blinded method. The interim analysis following the blinded adaptation also effectively limited the sample size when a strongly positive interim result was observed. The blinded sample size re-estimation used was transparent and could be used simply by the sponsor and investigators in a non-controversial way (Center for Biologics Evaluation and Research and Center for Drug Evaluation and Research, 2019). The positive interim results from this trial and the CAPTURE trial that read out at essentially the same time were a critical boost for the sponsor at that time.

EPILOG lessons learned included:

- Blinded sample size re-estimation in combination with group sequential design can be an effective combination to right-size a trial and come to a definitive evaluation early with strong positive results or late if the sample size adaptation is needed to ensure adequate events.
- While an early analysis had not shown differences in the previous EPIC trial, improvements in treatment regimens appears to have made enough difference to enable a definitive early analysis in the EPILOG trial.
- Multiple experimental arms were useful to best evaluate the risk-benefit of alternative heparin regimens.

5 CAPTURE

The CAPTURE Investigators et al. (1997) studied participants with refractory unstable angina undergoing PTCA who were at high risk for recurrent events. Treatment included medical therapy starting 18–24 hours prior to planned PTCA through 1 hour post PTCA. The adjudicated primary 30-day efficacy endpoint was similar to the EPIC trial primary endpoint above. The CAPTURE design was a double-blind randomized comparison of abciximab bolus and abciximab infusion (experimental) versus placebo bolus and placebo infusion (control). Proof of concept for safety and efficacy of the experimental treatment was demonstrated in randomized 60 patient Phase 2 study (Simoons et al., 1994). A group sequential design with 1400 participants was planned in CAPTURE to detect a reduction in the primary endpoint from 15% to 10% with 80% power and 2-sided $\alpha = 0.05$; we use one-sided testing here and convert study bounds appropriately. A time-to-event analysis using the logrank test was specified; this had little impact on sample size and power compared to a binary outcome in this case with a 30-day outcome and complete follow-up on essentially all participants. Interim analyses were planned after 25% (N=350) and 50% (N=700) of the 1400 planned participants were enrolled.

The group sequential design used a custom spending function to generate fit-for-purpose bounds as later outlined in Anderson and Clark (2010). This was inspired by discussion with Jan Tijssen, the DMC statistician from Amsterdam UMC. We present results here with one-sided testing as opposed to two-sided reporting by the CAPTURE Investigators et al. (1997). The spending function resulted in 1-sided bounds with $p = 0.00005$, $p = 0.0005$ nominal p-value bounds at 25% and 50% of participants to ensure that statistically significant differences would reflect clinically important differences that could justify a change in clinical practice. Following the second interim analysis, the DMC suggested an additional analysis after data from 75% of participants was available; the custom spending function gave a nominal 1-sided p-value of 0.0036; while the actual spending function was not published with the paper, these bounds are equivalent to a t-distribution spending function (Anderson and Clark, 2010) with cumulative spending of 0.00005, 0.000535, and 0.0038 at the 3 analyses; this can be implemented using Anderson (2020) or more simply with the Shiny interface at `https://rinpharma.shinyapps.io/gsdesign/`. While having an unblinded body recommend an additional analysis can inflate Type I error slightly even with spending functions, this is not generally a big issue with a smooth spending function (Proschan et al., 1992, Lan and DeMets (1989)). The addition of an interim analysis was discussed with regulators prior to implementing. Another fact that reduced any concerns about an observed p-value of 0.0032 1-sided at the N=1050 analysis, which is relatively close to the bound nominal p-value of 0.0036, was a strongly positive result in the related EPILOG trial shortly before this DMC review. As seen in Table 2, the interim bounds at analyses 2 and 3 are more conservative than the typical O'Brien-Fleming-like spending (Lan and DeMets, 1983); the corresponding comparison of spending functions is in Figure 2A. The chosen spending approach is consistent with requiring stringent interim results prior to stopping a trial at an interim analysis. The approach also enabled a less stringent result for a positive finding at the final analysis by spending less at interim analyses 2 and 3.

Table 2: CAPTURE bounds versus traditional O'Brien-Fleming-like spending bounds; nominal 1-sided p-values.

Spending	Interim 1	Interim 2	Interim 3	Final analysis
	(N=350)	(N=700)	(N=1050)	(N=1400)
Custom spending	0.000050	0.0005	0.0036	0.0244
O'Brien-Fleming	0.000007	0.0015	0.0092	0.0220

Figure 2B shows several things on a B-value scale. At an analysis proportion t_i into the trial relative to the final planned sample size, the B-value is defined as $\sqrt{t_i}Z_i$ where Z_i is a standardized normal test. In this case

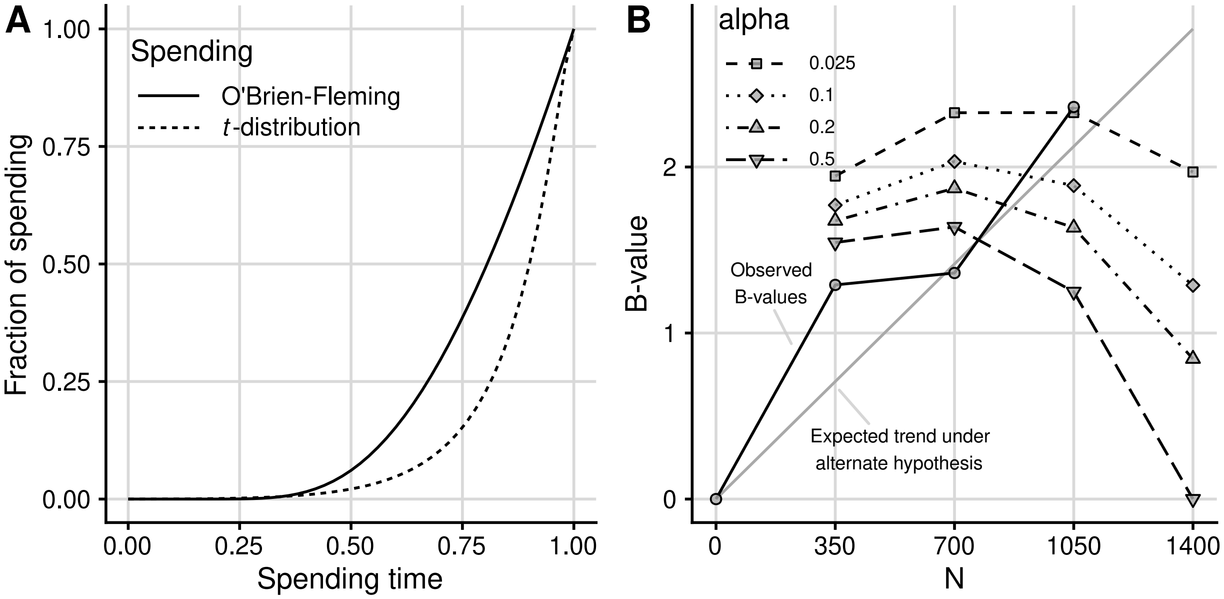

Figure 2: Panel A shows a comparison of O'Brien-Fleming-like spending functions and the custom spending function. Panel B shows the CAPTURE trial results and design sample space ordering.

$$Z_i = \frac{\bar{p}_{Ci} - \bar{p}_{Ei}}{\sqrt{\widehat{\text{Var}}(\bar{p}_{Ci} - \bar{p}_{Ei})}}. \tag{30}$$

That is, Z_i is the difference in observed event rates at analysis i divided by its estimated standard error. B-value scaling (Proschan et al., 2006) implies the expected value of each B-value is proportional to the number of observations as indicated by the diagonal line in the plot. The observed B-values are shown with circular points through the third analysis when the trial was stopped after crossing the interim efficacy bound. While the third interim and final results were based on the CAPTURE Investigators et al. (1997), the first two values were based on a binomial analysis of counts published in Cytel, Inc. (2007). The slope of the line from the origin to an interim observation represents the standardized effect size observed. Extending that line forward shows you what future predicted B-values would be under the same trend. For instance, the steep slope of the observed line from 0 to N=350 suggests that the trial would be expected to stop when N=700 if the same trend continued. However, we see the slope from N=350 to N=700 was almost flat, indicating little treatment effect in the second 350 patients. We see that the observed slope from 0 through the N=700 analysis is almost the same as the expected trend line indicated for the alternate hypothesis. We see also that extending that line would come very close to what was observed on the observed line at N=1050. The B-value at N=1050 crossed the group sequential bound for the trial indicated by the $\alpha = 0.025$ line. While the differences over time could just have been random variation, the countries contributing substantially to enrollment over time did change. While this possible lack of homogeneity in the patient population does not substantially impact a group sequential trial, it could have unfortunate implications for an adaptive trial design that depends heavily on homogeneity.

The $\alpha = 0.025, 0.1, 0.2$ and 0.5 lines in Figure 2B indicate that the entire set of outcomes can be ordered by group sequential bounds for different α-levels. This ordering uses the same spending function and timing of analyses with different α-levels to order the sample space. This ordering can be used to compute repeated p-values (Jennison and Turnbull, 1999) by selecting the smallest α-level at which a test rejects the null hypothesis; these can be considered multiplicity adjusted p-values. The minimum of repeated p-values up to and including an analysis is referred to as a sequential p-value by Liu and Anderson (2008), another form of adjusted p-value having the advantage that it cannot increase over time. Corresponding to repeated p-values are repeated confidence intervals (CI). Assuming a test can be performed of any null hypothesis real-valued θ_0, the lower repeated confidence interval is the smallest θ_0 value for which we can reject θ_0 in favor of a larger value. The convention used here, even for asymmetric testing, is to invert the test in the opposite direction at the same level. As noted by Jennison and Turnbull (1999), this results in a simultaneous coverage guarantee for all the intervals. Note that even after a bound has been crossed, if the original design is followed subsequent repeated p-values and confidence intervals maintain the above interpretability.

The result of the CAPTURE trial at N=1050 barely crossed the group sequential bound. Had the trial continued, the repeated and/or sequential p-values for the further planned analyses could still be legitimately interpreted as adjusted p-values.

We note that the DMC and sponsor agreed to stop enrolling in the trial at the time of a positive finding. This resulted in a total of 1266 patients reported in the primary publication, less than the final planned 1400 participants. The final 1-sided p-value with 1265 evaluable patients was 0.006. The inverse normal method of Lehmacher and Wassmer (1999) or a conditional error method such as Proschan and Hunsberger (1995) could be used to update bounds after such an adaptation is made when unblinded results are known. While the analysis for CAPTURE used the logrank test, only participants with complete 30-day data were included at each analysis. Thus, the independent increments related to participant groups added at each analysis held, enabling the standard asymptotic approach. A recent example in COVID-19 treatment with changes in treatment group differences early and late in the trial was Jayk Bernal et al. (2022).

We further characterize the results at interims 1–3 in the Table 3. Conditional power is evaluated using the interim treatment effect (CP thetahat) as well as the treatment effect for which the design was powered (CP theta1). Conditional error is the same calculation assuming $\theta = 0$. The strong and weak prior distributions used for Bayesian calculations are similar in nature to those suggested by Freedman and Spiegelhalter (1989). These work with the asymptotic distribution for the risk difference. Both prior means center on 0 risk difference, meaning that Bayesian analyses will shrink towards no difference compared to frequentist analyses. The strong and weak prior information are equivalent to about 10% and 1% of the information from the actual observations. The early very positive results are stronger than later in the trial. This reinforces the potential value of requiring very extreme results for an early stop for efficacy. The Bayesian analyses shrink early results towards no difference, suggesting the potential value of the predictive power evaluations and posterior mean differences shown in order to de-emphasize large early differences at the time of interim decision making. By the third interim when the trial was stopped, different evaluations of conditional and predictive power are similar as are the alternative effect size estimates. We note that the repeated confidence interval at interim 3 crosses 0, making it inconsistent with the statistically significant repeated p-value shown. This is because original data were not available to compute repeated confidence intervals with a Cox model consistent with the logrank test used for the final analysis. All other testing and confidence intervals in the table are based on the methods for unstratified risk difference of Miettinen and Nurminen (1985).

Table 3: CAPTURE evaluation at interim analyses.

	Interim analysis		
Measure	1	2	3
Sample size	350	700	1050
Control: Events/N	30/175	55/353	87/532
Control: Rate	17.1%	15.6%	16.4%
Treatment: Events/N	14/175	37/347	56/518
Treatment: Rate	8%	10.7%	10.8%
Rate difference	9.1%	4.9%	5.5%

Repeated CI	(-5.1%, 24%)	(-3.6%, 13.6%)	(-0.1%, 11.3%)
Z test statistic	2.58	1.93	2.73
Z-bound	3.89	3.29	2.69
p-value (1-sided, nominal)	0.005	0.027	0.003
Repeated p-value	NS	NS	0.022
CP (thetahat)	100%	85.8%	99.1%
CP (theta1)	95.4%	87.5%	98.6%
Conditional error	22.8%	19.9%	78.3%
Predictive power (strong prior)	83.7%	64.1%	96.2%
Predictive power (weak prior)	95.7%	76.1%	97.8%
Posterior mean (weak prior)	8.8%	4.8%	5.5%
95% prediction interval (weak prior)	(2%,15.6%)	(-0.1%,9.8%)	(1.3%,9.6%)
Posterior mean (strong prior)	6.5%	4.1%	4.8%
95% prediction interval (strong prior)	(0.6%,12.4%)	(-0.5%,8.6%)	(1%,8.7%)

CAPTURE lessons learned included:

- Enrolling in different sites over time can make trial non-homogeneous. While this is not a huge issue for group sequential trials, it could be problematic for trials with adaptive designs depending on interim trends.
- While O'Brien-Fleming-like spending is a good default, it may be worth considering a bespoke (custom) spending function.
- Using B-values and adjusted p-values on plots can be useful interim analysis tools.
- Bayesian analyses as well as conditional power based on different effect sizes can add useful perspective at the time of interim analysis to de-emphasize early extreme

differences.

6 Example testing many hypotheses

Graphical multiplicity control in group sequential design (Maurer and Bretz, 2013) is a powerful and flexible approach to evaluate multiple hypotheses in a group sequential design. An overview of graphical testing such as Bretz et al. (2011) may be useful for those not familiar with it. It is generally a simple way to communicate testing of multiple hypotheses and associated Type I error reallocation strategies. This was successfully taken to an extreme in the KEYNOTE-048 trial (Burtness et al., 2019) where 14 hypotheses were tested. The trial included 882 participants with untreated locally incurable recurrent or metastatic head and neck squamous cell carcinoma (HNSCC). Control treatment included cetuximab with chemotherapy. There were 2 experimental treatment groups: pembrolizumab monotherapy and pembrolizumab with chemotherapy. There were two primary outcomes: overall survival (OS) and progression free survival (PFS). Finally, there were 3 populations of interest based on combined positive score (CPS) for PD-L1 biomarker status at baseline: CPS ≥ 20, CPS ≥ 1, and the total population. This led to 12 superiority hypotheses for experimental treatment versus control: 2 endpoints × 2 experimental regimens × 3 populations. In addition, there were non-inferiority hypotheses in the overall population for overall survival for each of the experimental arms versus control, bringing the total hypotheses tested to 14. The graphical multiplicity plan is illustrated in Figure 3. Most Type I error was initially allocated to selected OS hypotheses. Less was allocated to PFS where more events were expected. The arrows in the figure indicate proportions of α-to be reallocated from rejected hypotheses to not yet rejected hypotheses; Bretz et al. (2011) provide a useful introduction to the Type I error reallocation algorithm. Adding to the complexities of this testing were

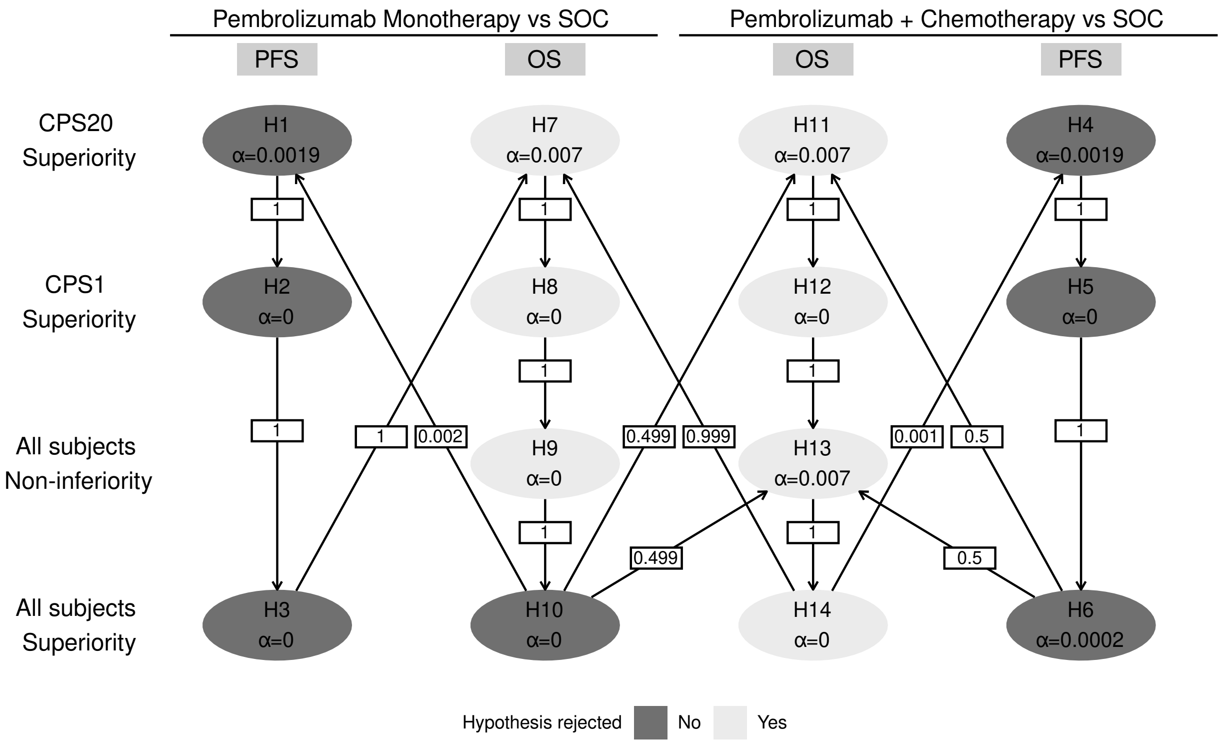

Figure 3: KEYNOTE-048 multiplicity strategy, based on Burtness et al. (2019).

multiplicity issues created by the two interim analyses planned.

6.1 Study results

Before going into details useful for study planning and execution, we note the study outcomes from Burtness et al. (2019) that demonstrate the value of the complex approach targeting multiple hypotheses. There were no statistically significant findings for the PFS endpoint for either experimental treatment compared to control in any of the pre-specified populations. However, overall survival for each experimental group was improved compared to control in both the CPS20 and CPS1 populations. For overall survival in the overall population the pembrolizumab + chemotherapy combination showed superior results and the pembrolizumab monotherapy group non-inferior results compared to control. At study start, it was not clear which of these hypotheses would likely be rejected; the α-splitting provided an opportunity to hedge bets for the study investment.

The α-splitting had the cost of more stringent bounds, but the benefit being if there were differential treatment effects by population, there could be a much better chance to find a positive result. The ability to define a benefiting population through multiple hypothesis testing in a single Phase 3 trial can be quite helpful to payers and prescribing physicians. Trying to find all these answers in a shorter, smaller Phase 2 trial can be a huge challenge. Adaptive designs possibly dropping some populations and reallocating α could have been problematic as the surrogates may have precluded the positive findings observed for pembrolizumab monotherapy.

6.2 Avoiding amendments

Burtness et al. (2019) was accompanied by supplementary materials that included the original and amendment 10 protocols. The final protocol indicated that some of the changes were related to challenges related to ensuring adequate event counts and follow-up duration for adequate evaluation of all study objectives. Without going through these changes in detail, we suggest the following general considerations that may be helpful in trials of this nature to reduce the need for protocol amendments. For any protocol, including this one, other considerations may drive specifications.

Changes to protocol hypotheses, if any, should be made early in the trial and prior to any unblinded analysis. When evaluating many hypotheses, it is extremely unlikely event targets for different hypotheses will be realized simultaneously. This suggests that the common practice of having all analysis timing be event-based is problematic in this type of trial. Differences in enrollment rates, subpopulation prevalence, control group event rates and treatment effects all interact to drive event accumulation for time-to-event hypotheses. A potential solution may be to focus on one population that is most important to get adequate data for. The sample size might focus on this population, suggesting that other populations will be studied with whatever sample size is available when the targeted population sample size is achieved. To illustrate this, we provide a simplified example that was not used for the actual protocol. For KEYNOTE-048, the original protocol suggested N=750 for the overall population and it was stated that the protocol may be amended if prevalence suggested fewer than 100 participants per arm would be enrolled in the CPS20 population (N=300 total). To avoid amendment, the protocol might be written to enroll 300 participants who were in the CPS20 population as well as up to 900 overall. The first two interim analyses were focused on PFS efficacy analysis. The second of these was the final planned analysis of PFS. Thus it may be useful to require both a minimum follow-up for the final participant enrolled as well as a PFS event target before the final cutoff for the analysis. This event target could be based on the CPS20 population with other populations and event counts being larger. Alternatively, timing could be based on a calendar duration of minimum follow-up since, in this metastatic indication, the disease recurrence rates in the control group were known to be very high. For instance, interim analyses could be planned with 6 and 12 months minimum follow-up regardless of events accrued. The final analysis was planned for OS only. Requiring a minimum follow-up (e.g, 2 years) could be useful to define the tail of the distribution for each population. Final analysis could require a targeted number of events for each population or simply cutoff at the targeted final follow-up duration. While a simple calendar cutoff could limit power if treatment effects are less than anticipated, it would ensure an appropriate assessment of tail behavior and complete the trial in a time frame that is relevant given the fast pace of changing practice standards for many cancers. The above considerations should provide adequate follow-up and reduce the need for amendments in this type of complex scenario.

Common spending function approaches (Maurer and Bretz, 2013) need some modification to incorporate the time-based analysis timing strategy above. Setting an incremental proportion of allocated α at each interim analysis (e.g., 1/25th) would work to appropriately order bounds as suggested by Liu and Anderson (2008). The final analysis then spends any remaining α; this is basically the modified Haybittle-Peto method suggested in Section 2.

Burtness et al. (2019) used the convention that the first analysis at which a boundary is crossed is the definitive analysis for a given hypothesis, treating subsequent analyses of the same hypotheses as supportive. This convention computes a study p-value for each hypothesis at the analysis where the result is first positive or, if never positive, at the final analysis performed. We will refer to this as the *first boundary crossed* convention. Thus, the study reports that at the second interim analysis H7, H8 and H9 were rejected, demonstrating improved survival for pembrolizumab alone versus control in the CPS20 or and CPS1 populations as well as non-inferior survival for the overall population. Also at the second interim analysis H13 and H14 were rejected demonstrating that overall survival was not only non-inferior but superior in the pembrolizumab + chemotherapy arm compared to control; at the final analysis H11 and H12 were rejected to definitively establish that pembrolizumab + chemotherapy was superior to control in the CPS20 and CPS1 populations. The p-values for these hypotheses at later analyses were reported, but not considered primary. The first boundary crossed convention used for this trial is not necessary since boundaries were chosen based on Maurer and Bretz (2013) which controls adjusted Type I error for all analyses performed per the trial design. That is, repeated p-values could be computed and reported as statistically significant due to the sample space ordering used to specifically enable reporting of repeated p-values or sequential p-values (Liu and Anderson, 2008) as discussed above with the CAPTURE study. The importance of this can be to add more definitive, longer-term evidence as equally and more important than evidence at the first time a boundary is crossed. In the case of Burtness et al. (2019), this could have meant greater emphasis in the primary manuscript on late treatment differences, given that more than 2 years of minimum follow-up was available by that time.

Other challenges are created by the approach here with multiple populations, dose groups, endpoints and interim analyses. First, it can be hard to predict subgroup prevalence. Second, both control and experimental event rates may vary between populations studied. For example, if prevalence of a subgroup were 40% instead of an assumption of 60%, accrual of planned events may be difficult or impossible in the subgroup. The group sequential design convention analyzing when a pre-defined fraction of final events is observed is essentially impossible when planned fractions will generally be realized at different calendar times for different hypotheses.

The implementation of Maurer and Bretz (2013) can be onerous; we have supported this with the R package **gMCPLite** (Zhu et al., 2022). The use of sequential p-values with the Maurer and Bretz method largely automated and documented which hypotheses could be rejected at each analysis. Pre-specifying how to deal with such logistical challenges is an important protocol consideration to avoid protocol amendments. The correlations between tests of different hypothesis are worth consideration; see Anderson et al. (2022b) for the implementation and accounting for correlations to relax bounds.

Lessons learned from the KEYNOTE-048 trial included:

- A large number of hypotheses can be successfully evaluated in a single trial to evaluate treatment strategy, population and multiple endpoints in a well-controlled fashion.

- Careful thought is required for timing of analyses as well as details of analysis to ensure the analysis is carried out in a rigorous way.
- The combination of open source software for design (**gsDesign**) and multiplicity (**gMCPLite**) are key to enabling this type of design.
- While preparation for this type of trial can be complex, underlying concepts of graphical testing and Type I error reallocation are straightforward to communicate with non-statisticians. The multiplicity strategy involves extensive discussions between statisticians and non-statisticians.

7 Discussion

We have found group sequential design to be a productive approach to the design of pivotal trials over many years. While the common use perhaps began within the National Heart Institute in the 1970s, group sequential trials have become quite common in the pharmaceutical industry and cooperative clinical trial groups since. With many standard textbooks and software, design and implementation can be straightforward. While there is also substantial software available for adaptive designs, challenges include fully accounting for operational challenges such as operational and medical practice changes over time. Basing changes on surrogate endpoints can also be problematic, as surrogacy assumptions may not be valid. Basing changes on early endpoints may be a mistake when the treatment difference between groups changes over time; on the other hand, group sequential design handles this naturally. There is often a desire to enroll trials quickly; while this can create challenges for group sequential trials, it can be particularly problematic for adaptive trials.

We have noted that logistics are key for any trials with interim analyses in order to ensure the best data for decision making. It is important to ensure that there are sufficient endpoints to do a meaningful analysis at early interims. It may be that the earliest interim analyses focus on harm rather than benefit, with stringent interim bounds usually recommended. Ensuring there is some time requirement as opposed to just event or patient counts between interim analyses can be worthwhile as fast information accrual can sometimes lead to small time gaps between analyses. For multi-population or multi-arm studies, carefully considering how boundaries will adapt when design assumptions are not met is important. In some cases with improving medical practice or a different patient population than expected, it can take a long time to accumulate targeted endpoints for all hypotheses being tested. In such cases, setting a maximum trial duration may be worthwhile even if this limits power.

In summary, group sequential design can right size a trial, allow evaluation of the balance of safety and efficacy over time and provide the best possible answers to difficult dose, population, and endpoint questions related to best medical practice. Finally, we have noted that interim bounds for efficacy and futility are worth considerable thought to ensure appropriate decision-making guidance.

References

Anderson, K., Zhang, Y., Zhao, Y., Xiao, N., Shirazi, A., and Yang, J. (2022a). *gsDesign2: Group Sequential Design with Non-Constant Effect.* R package version 1.0.0.

Anderson, K. M. (2020). *gsDesign: Group Sequential Design.* R package version 3.1.1.

Anderson, K. M. and Clark, J. B. (2010). Fitting spending functions. *Statistics in Medicine*, 29(3):321–327.

Anderson, K. M., Guo, Z., Zhao, J., and Sun, L. Z. (2022b). A unified framework for weighted parametric group sequential design. *Biometrical Journal*, 64(7):1219–1239.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal*, 53(6):894–913.

Burtness, B., Harrington, K. J., Greil, R., Soulières, D., Tahara, M., de Castro Jr, G., Psyrri, A., Basté, N., Neupane, P., Bratland, Å., et al. (2019). Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): a randomised, open-label, phase 3 study. *The Lancet*, 394(10212):1915–1928.

CAPTURE Investigators et al. (1997). Randomized placebo-controlled trial of abciximab before and during coronary intervention in refractory angina: the CAPTURE study. *The Lancet*, 349:1429–1435.

Center for Biologics Evaluation and Research and Center for Drug Evaluation and Research (2019). Adaptive designs for clinical trials of drugs and biologics guidance for industry. Technical report, United States Food and Drug Administration.

Chen, Y. J., DeMets, D. L., and Lan, K. K. G. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*, 23(7):1023–1038.

Cytel, Inc. (2007). EAST 5.

Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. (2019). *Data monitoring committees in clinical trials: a practical perspective*. John Wiley & Sons.

Ellenberg, S. S. and Shaw, P. A. (2022). Early termination of clinical trials for futility—considerations for a data and safety monitoring board. *NEJM Evidence*, 1(7):EVIDctw2100020.

Emerson, S. S., Kittelson, J. M., and Gillen, D. L. (2007). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*, 26(28):5047–5080.

EPIC Investigators (1994). Use of a monoclonal antibody directed against the platelet glycoprotein IIb/IIIa receptor in high-risk coronary angioplasty. *New England Journal of Medicine*, 330(14):956–961.

EPILOG Investigators (1997). Platelet glycoprotein IIb/IIIa receptor blockade and low-dose heparin during percutaneous coronary revascularization. *New England Journal of Medicine*, 336(24):1689–1697.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*. John Wiley & Sons.

Freedman, L. S. and Spiegelhalter, D. J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials*, 10(4):357–367.

Gandhi, L., Rodríguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., Domine, M., Clingan, P., Hochmair, M. J., Powell, S. F., et al. (2018). Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *New England Journal of Medicine*, 378(22):2078–2092.

Gerber, F. and Gsponer, T. (2016). gsbDesign: an R package for evaluating the operating characteristics of a group sequential Bayesian design. *Journal of Statistical Software*, 69:1–23.

Haybittle, J. (1971). Repeated assessment of results in clinical trials of cancer treatment. *The British Journal of Radiology*, 44(526):793–797.

Hwang, I. K., Shih, W. J., and De Cani, J. S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9(12):1439–1445.

Jayk Bernal, A., Gomes da Silva, M. M., Musungaie, D. B., Kovalchuk, E., Gonzalez, A., Delos Reyes, V., Martín-Quirós, A., Caraco, Y., Williams-Diaz, A., Brown, M. L., et al. (2022). Molnupiravir for oral treatment of Covid-19 in nonhospitalized patients. *New England Journal of Medicine*, 386(6):509–520.

Jennison, C. and Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, 74(1):149–154.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Lan, K. K. G. and DeMets, D. L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics*, 45(3):1017–1020.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Liu, Q. and Anderson, K. M. (2008). On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association*, 103(484):1621–1630.

Liu, Q. and Chi, G. Y. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics*, 57(1):172–177.

Magirr, D. (2021). Non-proportional hazards in immuno-oncology: Is an old perspective needed? *Pharmaceutical Statistics*, 20(3):512–527.

Magirr, D. and Burman, C.-F. (2019). Modestly weighted logrank tests. *Statistics in Medicine*, 38(20):3782–3790.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700.

Maurer, W. and Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research*, 5(4):311–320.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in medicine*, 4(2):213–226.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–891.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42(1-2):19–35.

Pampallona, S., Tsiatis, A. A., and Kim, K. (2001). Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal*, 35(4):1113–1121.

Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer*, 34(6):585–612.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Powles, T., Plimack, E. R., Soulières, D., Waddell, T., Stus, V., Gafanov, R., Nosov, D., Pouliot, F., Melichar, B., Vynnychenko, I., et al. (2020). Pembrolizumab plus axitinib versus sunitinib monotherapy as first-line treatment of advanced renal cell carcinoma (KEYNOTE-426): extended follow-up from a randomised, open-label, phase 3 trial. *The Lancet Oncology*, 21(12):1563–1573.

Proschan, M. A., Follmann, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48(4):1131–1143.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, pages 1315–1324.

Proschan, M. A., Lan, K. K. G., and Wittes, J. T. (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.

Roychoudhury, S., Anderson, K. M., Ye, J., and Mukhopadhyay, P. (2021). Robust design and analysis of clinical trials with nonproportional hazards: a straw man guidance from a cross-pharma working group. *Statistics in Biopharmaceutical Research*, pages 1–15.

Scharfstein, D. O., Tsiatis, A. A., and Robins, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association*, 92(440):1342–1350.

Simoons, M. L., de Boer, M. J., van Den Brand, M., Van Miltenburg, A., Hoorntje, J., Heyndrickx, G. R., van der Wieken, L. R., De Bono, D., Rutsch, W., and Schaible, T. F. (1994). Randomized trial of a GPIIb/IIIa platelet receptor blocker in refractory unstable angina. European Cooperative Study Group. *Circulation*, 89(2):596–603.

Slud, E. and Wei, L. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association*, 77(380):862–868.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, pages 193–199.

Wassmer, G. and Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*, volume 301. Springer.

Xi, D. and Gallo, P. (2019). An additive boundary for group sequential designs with connection to conditional error. *Statistics in Medicine*, 38(23):4656–4669.

Zhu, Y., Zhang, Y., Deng, X., Anderson, K., and Xiao, N. (2022). *gMCPLite: Lightweight Graph Based Multiple Comparison Procedures*. R package version 0.1.2.