Contemporary Clinical Trials 36 (2013) 587–596

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials

Wei Zhong [a], Joseph S. Koopmeiners [b], Bradley P. Carlin [b,*]

[a] Department of Biostatistics, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, United States
[b] Division of Biostatistics, MMC 303, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Received 24 October 2012
Received in revised form 16 March 2013
Accepted 20 March 2013
Available online 11 April 2013

Keywords:
Bayesian design
Clinical trial
Personalized medicine
Predictive approach
Sample size reestimation
Subgroup analysis

ABSTRACT

Frequentist sample size determination for binary outcome data in a two-arm clinical trial requires initial guesses of the event probabilities for the two treatments. Misspecification of these event rates may lead to a poor estimate of the necessary sample size. In contrast, the Bayesian approach that considers the treatment effect to be random variable having some distribution may offer a better, more flexible approach. The Bayesian sample size proposed by (Whitehead et al., 2008 [27]) for exploratory studies on efficacy justifies the acceptable minimum sample size by a "conclusiveness" condition. In this work, we introduce a new two-stage Bayesian design with sample size reestimation at the interim stage. Our design inherits the properties of good interpretation and easy implementation from Whitehead et al. (2008) [27], generalizes their method to a two-sample setting, and uses a fully Bayesian predictive approach to reduce an overly large initial sample size when necessary. Moreover, our design can be extended to allow patient level covariates via logistic regression, now adjusting sample size within each subgroup based on interim analyses. We illustrate the benefits of our approach with a design in non-Hodgkin lymphoma with a simple binary covariate (patient gender), offering an initial step toward within-trial personalized medicine.

1. Introduction

Traditional sample size determination for binary outcome data in a frequentist approach is simple, straightforward, and has been implemented in many clinical trials. For example, consider a two-arm clinical trial that compares the effect of two treatments, where we are interested in testing the hypotheses $H_0 : p_1 = p_2$ versus $H_a : p_1 > p_2$, where p_1 and p_2 denote the true event rates in the two treatment groups. To obtain the sample size given some pre-specified significance level α and power β, we must first set some target point estimates of p_1 and p_2 as crude guesses of the event probabilities for two treatments, denoting them as p_1^* and p_2^*, respectively. The designed detectable effect is then $\theta^* = p_1^* - p_2^*$. The sample size can be calculated by the following standard formula [15],

$$n \text{ per group} = \frac{2\left(Z_{1-\alpha/2} + Z_\beta\right)^2 \overline{p}^*(1-\overline{p}^*)}{\theta^{*2}}, \quad (1)$$

where the average event rate $\overline{p}^* = (p_1^* + p_2^*)/2$, and Z_γ denotes the γ percentile of the standard normal distribution.

Since the selection of p_1^* and p_2^* are usually based on fairly vague prior knowledge or other studies with small sample sizes, the credibility of the "working alternative hypothesis" that $p_1 = p_1^*$ and $p_2 = p_2^*$ is often questionable [24]. Misspecification of the event rates may lead to a poor estimate of the necessary sample size [22]. To fix this problem, many sequential designs and adaptive sample size designs incorporating interim analyses have been proposed in recent years [5,7,9,10,13,23]. All these methods can provide substantial

* Corresponding author. Tel.: +1 612 624 6646; fax: +1 612 626 0660.
E-mail address: brad@biostat.umn.edu (B.P. Carlin).

1551-7144/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cct.2013.03.011

improvement by adjusting the sample size to achieve the target power while preserving the overall Type I error. However, previous sample size reestimation methods are based on an implicit assumption that estimates of the true unknown treatment effect do not change appreciably over time. In real life situations, this assumption is questionable, especially when more subject-level variability exists in the early recruitment period. A good specification of the expected treatment effect is still required for these frequentist designs.

In contrast, the Bayesian approach considers the treatment effect to be random variable having some distribution, and updates the prior with the data, obtaining a posterior distribution for inference. The interpretation of a credible interval for the treatment effect seems more natural here than that of the traditional frequentist confidence interval. Moreover, the objective of a phase II trial is to accept or reject a new drug for further investigation in a phase III trial, rather than obtain a highly precise estimate of each possible response rate. Generally there are three classes of Bayesian methods for sample size determination. The first is a frequentist-Bayesian hybrid approach [2,16,17,25], which considers the predictive probability of achieving the primary study goal based on the available data, but still aims to control type I error. Second, some Bayesians recommend an interval length-based approach [14,18,19], which uses the length of posterior credible intervals as the sample size criterion. Finally, some authors pursue a fully decision-theoretic approach [1,4,21,26], which chooses sample size to maximize an investigator-selected utility function or minimize a corresponding loss function.

The Bayesian sample size proposed by [28] for exploratory studies on efficacy is an interval length-based approach, but includes an analogy to frequentist Type I and II errors. These authors argue that "the trial should be large enough to ensure that the data collected will provide convincing evidence either that an experimental treatment is better than a control or that it fails to improve upon control by some clinically relevant difference." Like frequentist designs, the expected treatment effect is explicitly set in the design. But Whitehead et al. sample size does not aim to meet certain power criteria under the alternative hypothesis. Instead, the acceptable minimum sample size N is justified by a "conclusiveness" condition. In the context of a one-sample test for a binary outcome (say, efficacy), it specifies that, regardless of the data, at least one of the two following probability statements should be satisfied at the end of a trial:

$$Pr\left(p > 0 \middle| Y^N\right) \geq \eta_1 \text{ or } Pr\left(p < \theta^* \middle| Y^N\right) \geq \eta_2, \tag{2}$$

where $p \in [0,1]$ denotes the success rate for the treatment, $\theta^* \in [0,1]$ is the expected (or desired) treatment effect, and Y^N represents any possible dataset of N patients. The threshold probabilities η_1 and η_2 are selected to reflect the degree of certainty we require for convincing evidence, with both values typically close to 1.

One potential problem is that such a sample size might be too conservative. Adding an interim stage to reestimate the sample size might offer a solution, dramatically reducing the sample size where the interim information about the true treatment effect emerges as sufficiently conclusive. Moreover, the corresponding Bayesian approach for comparing two proportions is not discussed by [28] and merits further exploration.

At the interim stage, one can calculate the predictive power based on the interim posterior estimates of the parameters. The predictive power is actually the "re-estimated" power based on the prior and the data. Thus, a Bayesian approach to sample size estimation seems more sensible and natural here. However, in contrast to the frequentist literature, sample size reestimation has been infrequently discussed in the Bayesian setting. Some Bayesians argue that Bayesian analysis is a naturally sequential procedure, and are thus unconcerned about Type I error inflation resulting from multiple interim looks. Patient recruitment should depend on the data available at that time, and the adequacy of the resulting predictive power for making a final decision. However in practice, the sample size is usually determined before starting the trial and the schedule of interim analyses is also fixed; many trialists feel it is inappropriate to adjust the recruitment plan during the trial. Sample size reestimation, a key factor in interim analysis, is thus relevant in Bayesian design as well. [27] applies a Bayesian predictive approach to interim sample size reestimation, and compares it to other approaches such as predictive and conditional power approaches. The author recommends its application in exploratory studies, where knowledge about a test drug is still uncertain, and the adaptive sample size is based on the predictive probability of trial success.

"Personalized medicine" is a subject of intense discussion in recent years. The concept refers to the tailoring of treatments to individuals based on personal characteristics, and represents the next step in drug therapy and development toward better understanding of disease and health [29]. The field is closely related to subgroup analysis, a subject of longstanding interest to trialists. For example, a recent study suggested no improvement in the overall mortality of patients with coronary disease whether treated with percutaneous coronary intervention (PCI) or coronary-artery bypass surgery. But the results also showed that age played a key role, with much lower mortality after surgery among patients 65 years or older, while lower mortality after PCI among those 55 years or younger [11]. Although many observational studies and pooled trials have contributed to our understanding of treatment effects at the individual level through subgroup effect analyses and development of prediction rules, a significant obstacle to the implementation of a personalized approach to trials themselves is the lack of appropriately designed studies [8]. Sample size estimation is an important issue for adequate trial design when we seek to study subgroup effects, especially in view of the well-known risk of Type I error inflation resulting from subgroups chosen post-hoc.

In this work, we first introduce a new two-stage Bayesian two-arm phase II trial design with sample size reestimation by implementing a predictive approach using [28] stopping rule. Then, we extend it to a four-subgroup trial design that considers an important binary covariate (gender) crossed with the treatment effect. Bayesian methods offer a direct attack on this problem, providing probabilities of efficacy, futility, and the like given the data seen so far. Traditional frequentist tools do not do this, and in fact p-values tend to overstate the evidence against H_0; worse, they are often