Published in final edited form as:

Curr Opin Oncol. 2008 July ; 20(4): 407–411. doi:10.1097/CCO.0b013e328302163c.

Bayesian designs to account for patient heterogeneity in phase II clinical trials

Peter F. Thall, PhD and J. Kyle Wathen, PhD

Department of Biostatistics, Box 447, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA

Abstract

Purpose of review—Between-patient heterogeneity is very common in clinical trials. This complicates treatment evaluation, due to known prognostic subgroup effects or potential treatment-subgroup interactions. We review two Bayesian phase II clinical trial designs that account explicitly for patient heterogeneity. The first design uses analysis of covariance to assess treatment effects in subgroups known to have different prognoses. The second design uses a hierarchical model for settings where, *a priori*, the experimental treatment effects in the subgroups are assumed to be exchangeable.

Recent findings—Compared to simpler designs that ignore patient heterogeneity, each design provides substantial improvements by reducing both false positive and false negative rates and focusing resources on subgroups where an experimental treatment is more likely to provide an advance over standard therapy. In either case, accounting for potential treatment-subgroup interactions is extremely important.

Summary—Due to the rapidly increasing number of potential new treatments to be evaluated clinically, increasing costs and limited resources, it is critically important to perform early phase clinical trials efficiently. The new Bayesian methods described here, and other related methods, provide efficient, broadly applicable tools to address these problems.

Keywords

Analysis of covariance; Bayesian design; hierarchical model; phase II clinical trial

Introduction

Although between-patient heterogeneity is common in clinical trials, most phase II designs assume patients are homogeneous [1–6]. This may lead to severe errors when evaluating an experimental treatment, E. A design that ignores known differences between prognostic subgroups runs substantial risks of reaching erroneous conclusions about the effects of E within subgroups. Alternatively if, *a priori*, there is no known difference between subgroups, assuming one overall response rate does not accommodate the possibility that the

Correspondence to Peter F. Thall, telephone 713-794-4162, rex@mdanderson.org.

Conflict of Interest Statement

None declared

effectiveness of E may differ between subgroups. We will describe two Bayesian designs [7, 8] that address these settings by accounting for patient heterogeneity.

Some Bayesian concepts

Since both of the designs we will discuss are Bayesian, we first review some Bayesian concepts. The two main objects in a statistical analysis are the observable data and the model parameter, θ . Parameters are not observable but rather are conceptual objects that characterize important aspects of the phenomenon being studied, such as the probability of tumor response, median survival time, or effects of patient characteristics on clinical outcome. Often, θ is a vector including several such quantities. In the Bayesian paradigm [9–11] θ is considered random. To reflect one's knowledge and uncertainty about θ before an experiment, a probability distribution, $\text{prior}(\theta)$, is assumed. For example, when asked to specify the probability of tumor response with standard therapy, S, for a particular disease, an oncologist might reply "Somewhere between 16% and 35%." Denoting the response probability with S by θ_S , the assumption that θ_S follows a $\text{beta}(20, 60)$ prior distribution yields the equation $\Pr(.16 < \theta_S < .35) = .95$, represented by the shaded area under the solid curve in Figure 1A. That is, under this probability distribution, $[.16 - .35]$ is a Bayesian "95% credible interval" for θ_S , which formalizes the oncologist's statement. Equivalently, a $\text{beta}(20, 60)$ distribution corresponds to having previously observed 20 responders and 60 non-responders in 80 patients treated with S. Suppose that θ_E is the response probability with an experimental treatment, E, in a planned phase II trial. To reflect substantial uncertainty about E one might assume that θ_E follows a $\text{beta}(.25, .75)$ prior, represented by the dotted curve in Figure 1A. These two distributions have the same mean, .25, but the $\text{beta}(.25, .75)$ has effective sample size of 1 patient whereas the $\text{beta}(20, 60)$ has effective sample size of 80 patients, reflecting almost no knowledge about θ_E and substantial knowledge about θ_S .

Once data have been observed, Bayes Law is applied to compute a new probability distribution, $\text{posterior}(\theta \mid \text{data})$, that quantifies what has been learned about θ from the data. The vertical bar can be read to mean "given that one knows" or "given that one has observed." The posterior is used to make statistical inferences about θ . In contrast, classical "frequentist" statistics considers θ to be fixed but unknown. Advances in computational methods have greatly accelerated practical application of Bayesian methods in recent years, especially in biostatistics [12] and clinical trials [13–16]. The Bayesian paradigm provides a natural framework for making decisions sequentially based on accumulating data during a clinical trial, as is done in a phase II trial with interim monitoring rules, since Bayes' Law may be applied repeatedly by using the posterior obtained after each stage as the prior for the next stage. In the above example, a Bayesian rule to stop the trial of E due to futility might take the form $\Pr(\theta_S + .15 < \theta_E \mid \text{data}) < .05$, which says to stop accrual if a .15 improvement of E over S in response probability is unlikely, given the observed data. Figure 1B gives the posteriors of θ_E that would result if 3 out of 10 (30%), 8 out of 20 (40%), or 20 out of 40 (50%) responses were observed on E in a phase II trial. For these three outcomes, the respective values of the decision criterion probability $\Pr(\theta_S + .15 < \theta_E \mid \text{data})$ are .23, .47 and .85.

Phase II trials with known prognostic subgroup effects

Most designs for phase II trials make the simplifying assumption that a single response probability applies to all patients, i.e., that patients are homogenous. While technically convenient, this assumption may lead to highly flawed conclusions. We will discuss two cases. In the first case, there are known prognostic differences between patient subgroups. We will show that a statistical design that ignores prognosis by assuming one response probability, θ_E , with E for all patients has a substantial risk of producing false positive or false negative conclusions about the effects of E within subgroups.

For simplicity, we will illustrate the first design in the case of two known prognostic subgroups, good (G) and bad (B). The design may easily be applied more generally in settings with more than two subgroups, although the per-subgroup sample size must be large enough to obtain reasonably reliable results. In this example, to keep track of the four treatment-subgroup combinations, we will denote the subgroup-specific tumor response probabilities by

$$\begin{aligned}\theta_{S,G} &= \text{Prob}(\text{response} | \text{treated with S in subgroup G}), \\ \theta_{S,B} &= \text{Prob}(\text{response} | \text{treated with S in subgroup B}), \\ \theta_{E,G} &= \text{Prob}(\text{response} | \text{treated with E in subgroup G}), \\ \theta_{E,B} &= \text{Prob}(\text{response} | \text{treated with E in subgroup B}).\end{aligned}$$

The method of Wathen et al. [7] is based on a Bayesian analysis of covariance (ANCOVA) model for these response probabilities that incorporates the historical subgroup effects seen with S and also allows the subgroups to have different E-versus-S treatment effects. The model expresses each logit (θ) = $\log\{\theta/(1-\theta)\}$ as

$$[\text{treatment effect}] + [\text{subgroup effect}] + [\text{treatment} \times \text{subgroup interaction}].$$

For example, $\text{logit}(\theta_{E,G}) = [\text{treatment E effect}] + [\text{subgroup G effect}] + [E \times G \text{ interaction}]$. Because the four probabilities given above share common parameters and are correlated under this model, it allows one to “borrow strength” between subgroups. Using the Bayesian formalism, an informative prior on the [subgroup effect] parameters is derived from historical data on S, while non-informative priors are assumed for all other parameters since they involve as yet unknown effects of E. Essentially, the data from the trial are used to learn about parameters involving effects of E. The design allows the possibility that E may be more effective than S in one subgroup but equivalent or inferior to S in another. For example, if an improvement of .15 in response probability is targeted within each subgroup then the criterion probability $\Pr(\theta_{S,G} + .15 < \theta_{E,G} | \text{data})$ is used to decide whether to stop or continue in the G subgroup and, similarly, $\Pr(\theta_{S,B} + .15 < \theta_{E,B} | \text{data})$ is used in the B subgroup. That is, the method uses subgroup-specific futility stopping rules, with the important provision that each rule is constructed so that the false negative rate (FNR) each subgroup is controlled below a specified small value, such as .10. Given a maximum total sample size, M, if accrual to the B subgroup is stopped early but the G subgroup is not

stopped, then accrual of G patients is continued until a total of M patients have been enrolled in the trial. In this way, the design spends limited resources in subgroups where E is more likely to provide an advance over S.

As an example, suppose it is known that, with S, the probability of tumor response is .45 in patients with good prognosis and .25 in patients with bad prognosis. The most common approach that accounts for prognosis is to simply conduct two separate trials, one in G patients and the other in B patients. Using either this approach or the Wathen et al. ANCOVA-based design [7], a response rate of $.45 + .15 = .60$ is considered “promising” in the G subgroup whereas .45 is not promising. Similarly, a response rate of $.25 + .15 = .40$ is considered “promising” in the B subgroup whereas .25 is not promising. A third approach, which is the most common, is to ignore prognosis by assuming one overall response rate with S, the average value $\theta_S = .35$, and one corresponding response rate θ_E with E. For a statistical design based on these simplifying assumptions that targets an improvement of .15 over S, $\theta_E = .50$ corresponds to E being considered promising whereas $\theta_E = .35$ or smaller implies E is not promising.

Table 1 gives computer simulation results comparing these three designs for a 100 patient trial of E monitored in cohorts of size 10. In Case 1, the null case where E provides no improvement over S in either subgroup, all three designs have reasonably high early stopping probabilities. The design that ignores subgroups appears to be the best approach of the three because it has the highest probability of correctly rejecting E, .86. Equivalently, it has the smallest false positive rate, $FPR = 1 - .86 = .14$. The approach of conducting separate trials within subgroups has smaller probabilities of rejecting E than the ANCOVA-based design in each subgroup because it does not borrow strength between subgroups. In Case 2, where E achieves the desired the .15 improvement over S in both subgroups, all three designs control the FNP at .10. In Case 3, there is treatment-subgroup interaction since E provides the desired improvement in the B subgroup, from .25 to .40, but not in the G subgroup. As in Case 1, the design that conducts separate trials within subgroups has a lower probability of rejecting E in the G subgroup compared to the ANCOVA-based design. However, the design that ignores subgroups has extremely poor properties, since the probability of stopping early and rejecting E equals .41 in both subgroups. This translates into $FPR = 1 - .41 = .59$ in the G patient subgroup and $FNR = .41$ in the B subgroup. While these error rates may seem surprisingly large, they illustrate the severe risks one runs if known patient heterogeneity is ignored. The flaw in ignoring the prognostic subgroups in this example can be seen by noticing that, accounting for patient heterogeneity, since a response rate of .40 is promising for the B subgroup and a response rate of .60 is promising for the G subgroup, the single value .50 targeted by the simple design is actually too small for G patients and too large for B patients. Unfortunately, the simple but error-prone approach of ignoring prognostic subgroups is that most commonly used in phase II trials.

For comparability, all three of these designs considered above had maximum sample size 100 and monitoring schedule using cohorts of size 10. If one wishes to also consider the commonly used Simon optimal two-stage design [3], one must assume homogeneity (ignore subgroups) in order to apply this design. In a similar setting where there is a treatment-subgroup interaction, a Simon design with up to 43 patients and nominal FNR and FPR both

set equal to .10 may have an actual FNR as high as .75 (Wathen et al., [7, Table III]). This large error rate has two causes, the first being that known prognostic differences are ignored, as in the above example. The second cause arises from the fact that, in general, a two-stage design is far less efficient than a design that monitors the accruing data more frequently.

Phase II trials starting with exchangeable subgroup effects

The second setting is that where, *a priori*, there is no known prognostic difference between subgroups and it is desired to conduct one trial including all subgroups. To simplify the illustration, we will consider a phase IIA trial where there is no standard treatment with anti-disease activity in any subgroup. The usual approach is to target a small fixed positive response probability, p , usually in the range .10 to .30. A Gehan design [1] is often used, or a Bayesian rule [17] that stops the trial if $\Pr(\theta > p \mid \text{data})$ is below a fixed small cut-off. Suppose that one wishes to allow the possibility that the effectiveness of E may differ between subgroups, but one also desires a statistical method in which a response observed in one subgroup leads to a higher estimated response rate in the other subgroups, that is, one wishes to borrow strength between subgroups. Neither assuming one θ for all subgroups nor conducting separate trials, one in each subgroup, provides a solution to this problem.

The second design [8] achieves both of the above goals by assuming a Bayesian hierarchical model. As a simple illustration, we will assume that there are 4 subgroups. Under the hierarchical model, there are four subgroup-specific response probabilities, $\theta_1, \theta_2, \theta_3, \theta_4$, that are generated from a common prior, the parameters of which themselves have second level prior, known as a “hyper-prior,” which induces correlation among the four probabilities. Consequently, for example, data in subgroup 1 affect not only the posterior of θ_1 , but also the posteriors of θ_2, θ_3 , and θ_4 . With this design, each subgroup has its own stopping criterion, $\Pr(\theta_1 > p \mid \text{data})$ for subgroup 1, $\Pr(\theta_2 > p \mid \text{data})$ for subgroup 2, etc., where, importantly, the “data” in each posterior probability includes the data from all 4 subgroups.

To see how this works in practice, suppose that a response probability .30 is targeted in each subgroup, we stop accrual to subgroup j if $\Pr(\theta_j > .30 \mid \text{data}) < 0.05$ and the prior correlation among the response probabilities corresponds to $\Pr(\theta_1 > .30) = .45$ whereas $\Pr(\theta_2 > .30 \mid 2 \text{ responses in 6 pts. observed in subgroup 2}) = .51$ and $\Pr(\theta_1 > .30 \mid 2 \text{ responses in 6 pts. observed in subgroup 2}) = .48$. This is the type of information elicited from the oncologists planning the trial that is used to establish the hyperprior [8, section 4]. Table 2 illustrates how the decisions made conducting one trial under this hierarchical model may differ from those that would be made if, instead, 4 completely separate trials were conducted assuming an independent prior for each θ_j . In Case 1, under the hierarchical model the higher empirical response rates of .50, .67 and .83 observed in subgroups 2, 3 and 4 pull the posterior of θ_1 upward so that, despite the low empirical response rate of .17 in subgroup 1, it is not stopped early. In contrast, if trials are conducted independently in each subgroup using 4 separate, non-hierarchical models these data would cause subgroup 1 to be terminated. Case 2 shows the same effect but with very different sample sizes within subgroups. Case 3 shows how this might work in the opposite direction, with the poor results in subgroups 1, 2 and 3 pulling θ_4 downward.

Figure 2 illustrates the effects of the assumed correlation among the four probabilities for a hypothetical trial taken from Case 3, with dashed lines showing the posteriors that would be obtained if θ_1 , θ_2 , θ_3 , θ_4 , were assumed to be independent and solid lines the corresponding posteriors under the hierarchical model. Note that the posteriors for θ_1 and θ_2 are shrunk more because subgroups 1 and 2 have smaller observed sample sizes of 5 and 10, whereas the posteriors for θ_3 and θ_4 are shrunk less because their sample sizes of 40 and 20 are much larger. This illustrates how the prior correlation induced by the hierarchical model affects the posteriors, and also the fact that, as with any reasonable Bayesian model, the accumulating data eventually overcome any prior assumptions.

Conclusions

We have briefly reviewed two general Bayesian methods that account for heterogeneity in phase II. More complex, application-specific Bayesian methods that account for patient covariates using regression models have been described [18–20]. A general conclusion is that treatment-subgroup interactions may cause simple phase II methods to have extremely large false positive and false negative error rates within subgroups. Consequently, not only accounting for patient subgroups, but also accounting for the possibility of treatment-subgroup interactions appear to be very important. Alternatively, if an investigator feels strongly that subgroup effects and treatment-subgroup interactions are highly unlikely, then a conventional design may be used.

A frequentist, hypothesis test-based approach to phase II trials that accounts for heterogeneity also has been proposed [21], but this method does not allow the possibility of treatment-subgroup interactions and thus suffers from potentially very large error rates within subgroups.

We have not addressed the related but very different problem of identifying new prognostic subgroups, which usually is not feasible in phase II due to limited sample sizes [22].

Acknowledgments

This research was partially supported by NIH grant RO1 CA-83932.

References and recommended reading

Papers of particular interest have been highlighted as :

- of special interest
 - of outstanding interest
1. Gehan EA. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *J Chronic Diseases*. 1961; 13:346–353. [PubMed: 13704181]
 2. Fleming TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982; 38:143–151. [PubMed: 7082756]
 3. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
 4. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*. 1995; 51:1372–1383. [PubMed: 8589229]

5. Thall PF, Simon R, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat in Medicine*. 1995; 14:357–379.
6. Rosner GL, Stadler W, Ratain MJ. Randomized discontinuation design: application to cytostatic antineoplastic agents. *J Clin Oncol*. 2002; 20:4478–84. [PubMed: 12431972]
- 7••. Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. *Stat in Medicine*. In press This method is reviewed in the present paper.
8. Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LO, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Stat in Medicine*. 2003; 22:763–780.
9. Gelman, A., Carlin, JB., Stern, HSA., Rubin, DB. *Bayesian Data Analysis*. 2. Boca Raton: Chapman and Hall/CRC; 2004.
10. Congdon, P. *Bayesian Statistical Modelling*. New York: Wiley; 2001.
11. Bolstad, WM. *Introduction to Bayesian Statistics*. New York: John Wiley; 2004.
12. Berry, DA., Stangl, DK., editors. *Bayesian Biostatistics*. New York: Marcell Dekker; 1996.
13. Racine A, Grieve AP, Fluehler H, Smith AFM. Bayesian methods in practice: experiences in the pharmaceutical industry. *Applied Statist*. 1986; 35:93–150.
14. Berry DA. A case for Bayesianism in clinical trials (with discussion). *Stat in Medicine*. 1993; 12:1377–1404.
15. Thall PF. Bayesian clinical trial design in a cancer center. *Chance*. 2001; 14:23–28.
16. Spiegelhalter, DJ., David, J., Abrams, KR., Myles, JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: John Wiley & Sons; 2004.
17. Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat in Medicine*. 1998; 17:1563–1580.
18. Thall PF, Sung H-G, Estey EH. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *J American Statistical Assoc*. 2002; 97:29–39.
- 19•. Thall PF, Wooten LH, Shpall EJ. A geometric approach to comparing treatments for rapidly fatal diseases. *Biometrics*. 2006; 62:193–201. A geometric representation of a two-dimensional parameter characterizing the probability of response and the mean time to response is used to design a randomized phase II trial to compare treatments while accounting for patient heterogeneity. [PubMed: 16542246]
- 20•. Thall PF, Nguyen H, Estey EH. Patient-specific dose-finding based on bivariate outcomes and covariates. *Biometrics*. In press. A method is proposed for dose-finding in phase I/II clinical trials based on both efficacy and toxicity that accounts for between-patient heterogeneity.
21. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Stat in Medicine*. 2005; 17:2597–2611.
22. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Stat in Medicine*. 2004; 24:329–339.

Figure 1A

Figure 1B

Fig. 1.

Fig. 1A. Illustration of beta distributions. A beta(20,60) is shown by the solid line, with its 95% credible interval shown by the shaded region under the curve. A beta(.25, .75) distribution is shown by the dotted line.

Fig. 1B. Illustration of beta distributions. A $\text{beta}(.25, .75)$ prior is shown by the solid line, and the posterior distributions that would be obtained if 3/10, 8/20 or 20/40 responses were observed are given by dashed lines.

Fig. 2. Illustration of the within-subgroup posteriors from a hypothetical trial under either a simple model with four independent response probabilities (dashed lines) or under the hierarchical model (solid lines).

Table 1

Computer simulation results for a trial with up to 100 patients, monitoring in cohorts of size 10. Each design targets an improvement of .15 over each assumed null response probability, .25 in the bad (B) subgroup, .45 in the good (G) subgroup, or the single average null value .35 in the design that ignores subgroups.

		Pr(Reject E) (Average No. patients)			
Case	Prognostic Subgroup	True Pr(Response)	Account for Subgroups Using ANCOVA	Conduct Separate Trials Within Subgroups	Ignore Subgroups
1	Good	.45	0.76 (38)	0.65 (34)	0.86 (24)
	Bad	.25	0.78 (37)	0.65 (33)	0.86 (24)
2	Good	.60	0.10 (50)	0.10 (47)	0.10 (47)
	Bad	.40	0.10 (50)	0.10 (47)	0.10 (47)
3	Good	.45	0.72 (33)	0.65 (34)	0.41 (38)
	Bad	.40	0.10 (64)	0.10 (47)	0.41 (38)

Table 2

Illustration of decisions made by futility stopping rules with one trial conducted under the hierarchical model, compared to conducting separate trials within the four subgroups. Each trial targets a .20 response rate in each subgroup. Each observed data entry is [# responses]/[# patients evaluated].

Case		Subgroup			
		1	2	3	4
1	Observed Data	5/30	15/30	20/30	25/30
	Hierarchical Model	Continue	Continue	Continue	Continue
	Four Separate Trials	Stop	Continue	Continue	Continue
2	Observed Data	0/5	3/10	18/20	25/40
	Hierarchical Model	Continue	Continue	Continue	Continue
	Four Separate Trials	Stop	Continue	Continue	Continue
3	Observed Data	1/30	1/30	2/30	6/30
	Hierarchical Model	Stop	Stop	Stop	Stop
	Four Separate Trials	Stop	Stop	Stop	Continue