Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Predictive Probability of Success and the Assessment of Futility in Large Outcomes Trials

Benjamin Trzaskoma & Andreas Sashegyi

To cite this article: Benjamin Trzaskoma & Andreas Sashegyi (2007) Predictive Probability of Success and the Assessment of Futility in Large Outcomes Trials, Journal of Biopharmaceutical Statistics, 17:1, 45-63, DOI: 10.1080/10543400601001485

To link to this article: https://doi.org/10.1080/10543400601001485

Published online: 02 Feb 2007.

Journal of Biopharmaceutical Statistics, 17: 45–63, 2007

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400601001485

Taylor & Francis
Taylor & Francis Group

PREDICTIVE PROBABILITY OF SUCCESS AND THE ASSESSMENT OF FUTILITY IN LARGE OUTCOMES TRIALS

Benjamin Trzaskoma and Andreas Sashegyi
Eli Lilly and Company, Indianapolis, Indiana

We consider a class of futility rules based on a Bayesian approach for computing the predictive probability of success for large clinical trials, given a certain amount of observed data. This paper focuses on outcomes trials in particular, thus we are concerned with binary response variables. The proposed method determines the likelihood of observing a statistically significant treatment effect at the end of a study, conditional on the data observed at an interim time point and assuming that event rates governing future observations follow beta distributions. In particular, the prior distributions for the event rates of interest are updated based on the observed data at an interim time point, such that means and variances are intuitive functions of the data. Computational aspects will be discussed for the case in which event counts are functions of sample size and event rates only, and for situations in which they are functions of sample size, event rates, and exposure duration. We will discuss appropriate thresholds for declaring futility based on this approach, and the potential impact of overdispersion, a common phenomenon particularly in global outcomes trials.

Key Words: Binary data; Futility rules; Overdispersion; Predictive probability; Prior distribution.

1. INTRODUCTION

Large, long-term clinical trials require an enormous investment of resources, and are therefore associated with considerable risk for the sponsor. An interim analysis strategy is typically applied in such studies, not only to ensure the earliest possible detection of potential safety concerns, but possibly also to allow the sponsor to terminate the trial early under certain conditions if the evidence of treatment efficacy is sufficiently strong. A further important stopping rule which should be considered, especially in very costly trials, relates to futility: If the observed safety profile is acceptable but it seems highly unlikely that the trial will ever meet its efficacy objective, then the most prudent decision may be to cut losses, terminate the study, and channel resources into more promising directions.

We consider assessment of futility based on a Bayesian approach for computing the predictive probability of success (PPS, sometimes referred to as "predictive power") for the trial, given a certain amount of observed data at an interim time point. PPS differs from conditional power (Halperin et al., 1982) in that the latter gives the probability that a statistically significant treatment difference will

Received July 11, 2005; Accepted June 16, 2006
Address correspondence to Dr. Andreas Sashegyi, Eli Lilly and Company, Lilly Corporate Center, DC 6072, Indianapolis, IN 46285; E-mail: aisasheg@lilly.com

be observed at the end of the trial, given the data collected thus far but treating certain assumed treatment and placebo event rates as fixed for future observations; often the protocol-specified event rates are used in computing conditional power. In contrast, PPS allows for variation in future event rates, by allowing the event rate parameters to vary according to distributions whose means and variances are determined by the data already observed. This distinction is important in evaluating the likelihood of a successful trial, especially if observed event rates at a given time point differ significantly from protocol assumptions, or if there is considerable uncertainty about the true underlying event rates. We compute PPS in a Bayesian framework, using the conjugate prior beta distribution as a natural and mathematically convenient choice. Gelman et al. (1995) provide a useful guide to Bayesian approaches in general.

Numerous authors have discussed the use of predictive probabilities and other futility rules in various contexts. Typically, a data monitoring committee (DMC) will use such rules at interim analyses, in conjunction with other guidelines, to form recommendations as to whether or not a study should be continued. Wittes (2000) provides a helpful introduction to DMCs. Jennison and Turnbull (2000) describe formal statistical methods for interim monitoring, ensuring that Type I and II error rates do not exceed pre-specified values. Early work of Herson (1979) specifically addressing the question of futility considered Bayesian rules for relatively small single-arm Phase II trials with dichotomous outcomes. Useful graphical summaries of predictive distributions that give a visual depiction of PPS are shown in Spiegelhalter et al. (1986). Others (Freidlin and Korn, 2002; Van der Tweel and van Noord, 2003) have discussed limitations and further comments on futility monitoring in general.

In Section 2, we contrast the formulation of PPS with conditional power, which is a special case of the former. We will also discuss the beta distribution as the natural choice in terms of the prior distribution for computing PPS in the binary response setting. The formulation of PPS presented in this paper is similar to that given in Johns and Andersen (1999), but the computational approach we describe specifically addresses large outcome trials in which the calculation of the probability of specific future outcomes, as these authors describe, is infeasible; Section 2 provides more detail in this regard. Hence, by referring to "large" outcomes trials we mean any study in which predictive probabilities need to be evaluated directly as opposed to summing probabilities of individual possible outcomes. Section 3 will discuss these computational aspects for the case in which event counts are functions of sample size and event rates only, and for situations in which they are functions of sample size, event rates, and exposure duration, which has not been addressed by the previous authors. We will provide examples of recent global outcomes trials that fall into these categories. In Section 4 we then examine the performance of predictive as compared to conditional power, and explore characteristics of PPS with respect to the choice of appropriate thresholds for declaring futility. Section 5 will address the potential impact of overdispersion on the computation of predictive probablilities. This phenomenon, also known as extra-binomial variation, is common particularly in global outcomes trials. Its impact on inferences can be dramatic, hence it is important to understand the circumstances under which overdispersion can be a concern. Further discussion is provided in Section 6.

2. PREDICTIVE PROBABILITY OF SUCCESS FOR LARGE OUTCOMES TRIALS

Suppose that test treatment T is being compared to control C with respect to a binary outcome such as mortality. Of interest in a clinical trial is the comparison of the event rates π_T and π_C in the two treatment groups, with the typical objective of testing the null hypothesis $H_0 : \pi_T - \pi_C \geq 0$ against the alternative $H_A : \pi_T - \pi_C < 0$. We will formulate expressions for PPS using the test statistic

$$Z = (p_T - p_C)/\sqrt{\text{Var}(p_T - p_C)} \tag{1}$$

in which p_T and p_C are estimates of π_T and π_C, and on the basis of which efficacy (success) is declared if $Z < -z_{\alpha/2}$, where z_q is the upper q percentile of the standard normal distribution. In general, letting r denote the number of events observed in N patients, we use the maximum likelihood estimates $(p_T, p_C) = (r_T/N_T, r_C/N_C)$. Test statistics other than Equation (1), such as the odds ratio or relative risk, could be used as well if this were advantageous. As long as appropriate critical values can be determined to differentiate a significant effect from a nonsignificant one, the complexity of the distribution of the test statistic itself does not increase the complexity of the computation of PPS, since this computation is easily achieved through simulation.

A special case of PPS is the probability of a successful outcome at the end of the trial, given certain observed data $\boldsymbol{x}_{\text{obs}} = (x_T/n_T, x_C/n_C)$ and assuming event rates π_T and π_C for future observations in the two treatment groups to be fixed quantities. This is simply the conditional power, defined as

$$CP = P(Z < -z_{\alpha/2} \mid \pi_T, \pi_C, \boldsymbol{x}_{\text{obs}}). \tag{2}$$

In this expression, Z is a function of the difference of observed final event rates as in Equation (1). Here p_T and p_C both have the form r/N, where $r = x + y$, in which x (out of n patients in a given treatment group) is the number of events already observed at the time CP is calculated, and y is the number of events occurring in the remaining $N - n$ patients. Thus, $\boldsymbol{y} = (y_T, y_C)$ is the only random component in Equation (2), Z can be written as $Z(\boldsymbol{y})$, and CP is evaluated assuming that (y_T, y_C) come from the distributions $\text{Bin}(N_T - n_T, \pi_T)$ and $\text{Bin}(N_C - n_C, \pi_C)$, respectively. Often the protocol-specified rates are used for π_T and π_C in Equation (2), but alternatives include placing emphasis exclusively on the observed data (i.e., conditioning not only on $\boldsymbol{x}_{\text{obs}}$ but also using the observed event rates for π_T and π_C) or using weighted averages of observed and assumed theoretical event rates for the future observations.

Equation (2) can be extended to a more general expression for PPS by allowing variation in the event rates assumed for future observations. This is also an intuitively appealing way of reconciling the fact that the observed data at an interim analysis may not be entirely consistent with the protocol-specified event rates. Moreover, it would be desirable for a futility rule to distinguish itself from conditional power not only by allowing future event rates to vary according to any reasonable distributions in general, but specifically by incorporating information from the observed data into the parameter assumptions about these distributions.

For example, it makes sense that as the amount of data observed increases, computation of PPS should use distributions for future event rates that are increasingly centered at the observed event rates, and whose variances approach zero. Such a general expression for PPS is given by the equation

$$\text{PPS} = \int \int P(Z < -z_{\alpha/2} \mid \pi_T, \pi_C, \boldsymbol{x}_{\text{obs}}) f(\pi_T, \pi_C \mid \boldsymbol{x}_{\text{obs}}) d\pi_T d\pi_C, \tag{3}$$

in which there is explicit dependence of the distribution of the event rates, $f(\pi_T, \pi_C \mid \boldsymbol{x}_{\text{obs}})$, on $\boldsymbol{x}_{\text{obs}}$.

It is instructive to contrast this expression for PPS with the analogous alternative typically used for smaller trials (Choi et al., 1985; Johns and Andersen, 1999). As above, let $\boldsymbol{x}_{\text{obs}} = (x_T/n_T, x_C/n_C)$ be the observed data at a given interim analysis, with $N_i - n_i$ future observations (patients) remaining in each of two treatment groups, $i = T, C$. The data yet to be observed are $\boldsymbol{y} = (y_T, y_C)$ where y_T and y_C can take on values from 0 to $N_T - n_T$ and $N_C - n_C$ respectively, according to the number of future events observed in each treatment group. PPS in this case is given by

$$\sum_{i \in \{0,\ldots,N_T - n_T\}} \sum_{j \in \{0,\ldots,N_C - n_C\}} \boldsymbol{I}(Z(\boldsymbol{y}) < -z_{\alpha/2} \mid \boldsymbol{x}_{\text{obs}}) P(y_T = i \mid \boldsymbol{x}_{\text{obs}}) P(y_C = j \mid \boldsymbol{x}_{\text{obs}}) \tag{4}$$

where $\boldsymbol{I}(\cdot)$ is an indicator function that takes value 1 if its argument is true and 0 otherwise, and

$$Z(\boldsymbol{y}) = ((y_T + x_T)/N_T - (y_C + x_C)/N_C)/\sqrt{\text{Var}(p_T - p_C)}. \tag{5}$$

The variance in the above expression can be estimated by the usual binomial expression $p_T(1 - p_T)/N_T + p_C(1 - p_C)/N_C$, where $p_T = (y_T + x_T)/N_T$ and $p_C = (y_C + x_C)/N_C$, or by using a pooled event rate estimate in place of p_T and p_C. Note that Equation (4) gives an expression for the proportion of times a successful result will be observed at the end of the trial, considering all $(N_T - n_T) \times (N_C - n_C)$ pairs of possible future outcomes, as well as the probability of each pair. If one assumes beta distributions for the true but unknown event rates of outcomes to be observed in future, which are updated at each interim analysis on the basis of the data already observed, it follows that $P(y_T = i \mid \boldsymbol{x}_{\text{obs}})$ and $P(y_C = j \mid \boldsymbol{x}_{\text{obs}})$ are beta-binomial distributions, which are the predictive distributions for future events. In small trials it is possible to evaluate predictive probabilities by computing probabilities for particular outcomes using expression (4). This becomes infeasible, however, in large trials where hundreds and sometimes thousands of observations remain to be collected at a given interim analysis. In such cases, PPS is computed by evaluating the integral in Equation (3) (see, for example Chapter 10 of Gelman et al., 1995). In the following section we discuss the details of implementing this approach in a computationally convenient manner by simulating future observations from the binomial distribution, but obtaining the binomial parameter from the posterior distribution of the event rate, that is, the beta distribution. A similar approach for obtaining a beta-binomial density using the Gibbs sampler is described in Casella and George (1992). PPS is discussed both for cases in which event counts are

functions of sample size and event rates only, and for situations in which they are functions of sample size, event rates, and exposure duration.

3. COMPUTATIONAL ASPECTS

Large outcomes trials may be characterized according to the duration of follow-up on patients, and the implications this has on the outcome of interest. In studies of acute conditions, patient follow-up at least for the primary endpoint is typically short, and exposure duration not a factor in the determination of outcome. In sepsis trials, for example (Abraham et al., 2005; Bernard et al., 2001), large numbers of patients may be enrolled and each observed for the outcome of all-cause mortality at a single landmark time point, 28 days after randomization. The ADDRESS trial (Abraham et al., 2005) (ADministration of DRotrecogin alfa (activated) in Early stage Severe Sepsis), which studied use of drotrecogin alfa (activated) in lower-risk severe sepsis patients is a particular example, which will be discussed in greater detail in Sections 4 and 6.

On the other hand, in studies of chronic conditions, patient follow-up may span years, with the likelihood of developing an event increasing with greater exposure. In cardiovascular and health outcomes studies (Hulley et al., 1988; Women's Health Initiative Investigators' Women's Health Initiative Investigators' Writing Group, 2002), patients are typically followed for several years, and thus event counts in such studies are not only functions of sample size and event rates per unit time, but also exposure duration. We discuss PPS for both types of trial designs, considering for the latter case endpoint-driven trials in which patient follow-up continues until the achievement of a fixed total number of endpoints. Motivation for evaluating PPS in this setting came from the recently completed Raloxifene Use in The Heart (RUTH) study (Barrett-Connor et al., 2006), investigating the effect of raloxifene in reducing the risk of major coronary events and invasive breast cancer in postmenopausal women.

All calculations were performed using S-PLUS® 6.2 for Windows, a convenient computing environment which supports straightforward random sample generation from various standard distributions such as the beta and the binomial.

3.1. PPS with Event Counts Functions of Sample Size and Event Rates Only

Assuming independence of event rates conditional on $\boldsymbol{x}_{\text{obs}}$, we have $f(\pi_T, \pi_C \mid \boldsymbol{x}_{\text{obs}}) = f_T(\pi_T \mid x_T) f_C(\pi_C \mid x_C)$. Before having observed any data, assume that event rate π_i, $i = T, C$ follows a beta distribution such that (dropping subscripts for simplicity)

$$f(\pi) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))\pi^{a-1}(1-\pi)^{b-1}, \quad a, b > 0.$$

Having observed x events (for example, deaths) in n patients in a given treatment group, and assuming that $x \mid \pi$ is distributed as $\text{Bin}(n, \pi)$, the posterior distribution of $\pi \mid x$ is

$$f(\pi \mid x) \propto p(x \mid \pi) f(\pi) \propto \pi^{x+a-1}(1-\pi)^{n-x+b-1} \sim \text{Beta}(a+x, b+n-x).$$

Hence, the posterior distribution of mortality rates is again a beta distribution, with mean and variance depending on the observed data and the initial parameters a and b. Note that

$$E(\pi \mid x) = (a + x)/(a + b + n) \quad \text{and}$$
$$\text{Var}(\pi \mid x) = E(\pi \mid x)(1 - E(\pi \mid x))/(1 + a + b + n).$$

As x and n grow large in relation to a and b, which happens quickly in large trials, the mean of the distribution for the future event rates approaches the observed event rates, and the variance approaches zero at a rate $O(n^{-1})$, an intuitively appealing feature. Choosing $a = b = 1$ initially is a reasonable starting point, corresponding to a uniform prior. PPS as given by Equation (3) can therefore be computed as follows:

1. Given $\boldsymbol{x}_{\text{obs}} = (x_T/n_T, x_C/n_C)$, draw a random pair of probabilities π_{T^*}, π_{C^*} from $f(\pi_T, \pi_C \mid \boldsymbol{x}_{\text{obs}})$.
2. Generate a pair of binomial random variables $\boldsymbol{y} = (y_T, y_C)$ from $\text{Bin}(N_T - n_T, \pi_{T^*})$ and $\text{Bin}(N_C - n_C, \pi_{C^*})$ and compute $Z = Z(\boldsymbol{y})$ (as in Equation [5]).
3. Repeat steps 1 and $2K$ times and determine PPS as the proportion of times that $Z < -z_{\alpha/2}$.

That is,

$$PPS \approx K^{-1} \sum_{i \in \{1,\ldots,K\}} \boldsymbol{I}(Z(\boldsymbol{y}) < -z_{\alpha/2} \mid \boldsymbol{x}_{\text{obs}}),$$

where $\boldsymbol{I}(.)$ is the indicator function as defined above. Note that computationally this algorithm combines two steps into one: instead of evaluating $P(Z < -z_{\alpha/2})$ for each of a large number of randomly selected probability pairs from the relevant beta distribution and then taking an average, a particular value of Z is computed once for each pair (π_{T^*}, π_{C^*0}), with $P(Z < -z_{\alpha/2})$ simply determined as the proportion of times $Z < -z_{\alpha/2}$. This approach is easily implemented and seems to work well in practice, with $K = 50,000$ being a sufficiently large number simulations to determine PPS to two decimal places of accuracy.

3.2. PPS with Event Counts Functions of Sample Size, Event Rates, and Exposure Duration

Consider now a trial design in which a cohort of patients is enrolled and followed up over time until a predetermined total number of events is observed. Such designs are common for example in cardiovascular and some oncology outcomes research. Final event rates π_T and π_C depend on time in this case, since increasing exposure duration implies increasing numbers of events. Suppose that within a particular treatment group, x events are observed among n patients at an interim analysis at which the average exposure duration per patient is M months. One cannot use the distribution of $\pi \mid x$ directly to model the number of events yet to be observed in the trial, since this depends on the final patient exposure duration, which itself is an unknown at the interim analysis. That is,

there is no distribution directly analogous to $\text{Bin}(N-n, \pi_*)$ in Step 2 above. It is reasonable to assume, nonetheless, that occurrence of future events will be governed by a *monthly* or otherwise appropriately standardized event rate that follows a distribution whose mean and variance can be estimated by the data observed. We estimate this distribution as

$$\pi_m \mid x \sim M^{-1}\pi \mid x \tag{6}$$

where $\pi \mid x \sim \text{Beta}(a+x, b+n-x)$ as above, and $a=b=1$ are indicated if one wishes to begin with a uniform prior distribution; samples from $\pi_m|x$ are simply generated by sampling from $\pi|x$ and dividing by M. In general, the mean patient exposure M can be expressed in any unit of time. This gives rise to the following algorithm for computing PPS:

Assume that x_T/n_T and x_C/n_C are the observed event rates ($\boldsymbol{x}_{\text{obs}}$) in two treatment groups at a given interim analysis.

1. Compute the mean exposure duration in each group, say M_T and M_C, expressing these quantities in terms of units of time that are small relative to the approximate follow-up time still expected in the trial.
2. Construct the standardized distributions $\pi_{Tm} \mid x_T$ and $\pi_{Cm} \mid x_C$ following Equation (6). (Assume for the purpose of illustration that a year or more of patient follow up is expected, and that M_T and M_C are expressed in months, so that the distributions generated describe monthly event rates).
3. Randomly sample $\pi_T{*}_m$ and $\pi_C{*}_m$, that is, one observation from each of $\pi_{Tm} \mid x_T$ and $\pi_{Cm} \mid x_C$.
4. Use these rates to compute the expected number of events yet to be observed in each treatment group in the trial, say y_T and y_C:

   a. Compute the expected number of events in each treatment group going forward, month by month. (At the first iteration the expected number of additional events will be $(n_T - x_T)\pi_T{*}_m$ and $(n_C - x_C)\pi_C{*}_m$).
   b. Keeping a running tally, after each month determine whether the total number of events exceeds the predetermined number specified for the trial.

      1. If it does, stop.
      2. If it does not, reduce the risk sets in the two treatment groups by the number of new patients with an event and proceed to calculations for the subsequent month.

5. Compute the projected final event rates $p_T = (y_T + x_T)/n_T$ and $p_C = (y_C + x_C)/n_C$, and the value of a test statistic of choice, $Z = Z(\boldsymbol{y}) = Z(p_T, p_C)$.
6. Repeat Steps 3 to $5K$ times and calculate, as in Section 3.1, $\text{PPS} \approx K^{-1} \sum_{i\in\{1,\ldots,K\}} \boldsymbol{I}(Z(\boldsymbol{y}) < -z_{\alpha/2} \mid \boldsymbol{x}_{\text{obs}})$.

This algorithm assumes that at the time of the interim analysis when PPS is evaluated, patient enrolment is complete so that n_T and n_C represent the final sample sizes of each treatment group. If enrolment were not complete, Step 4(a) would require appropriate modification to take the additional planned patient recruitment into account when computing the incremental expected number of events.

4. CHARACTERISTICS OF PREDICTIVE PROBABILITY OF SUCCESS

We now explore some of the properties of PPS by first considering specifics of the ADDRESS trial and generalizing from there through simulation of other relevant scenarios.

ADDRESS was a randomized, global placebo-controlled study to determine whether treatment with drotrecogin alfa (activated) reduces the risk of 28-day all-cause mortality in adults with early stage severe sepsis. Adjusting for 'drop-ins' (patients in the placebo arm receiving active treatment commercially), the trial assumed mortality rates in the placebo and active treatment arm of the study of 18.4% and 16%, respectively. This necessitated a sample size of approximately 11,400 patients to detect a significant treatment difference. Three interim analyses were planned for ADDRESS, after the enrolment of 1000, 3816, and 7632 patients, respectively. To place PPS in context for ADDRESS, Table 1 compares representative values of PPS from Equation (3) to analogous values of conditional power derived from Equation (2), assuming future event rates to be fixed at (i) the observed rates at each interim, (ii) the protocol-specified rates, and (iii) a weighted average of the two, with weights determined by the cumulative fraction of information collected at each interim. We show these comparisons for hypothetical calculations made at each of the three planned interim analyses of ADDRESS

Table 1 Conditional power and PPS for the ADDRESS trial

Approach	Interim 1 $n = 1000$	Interim 2 $n = 3816$	Interim 3 $n = 7632$
(A) Assumed mortality rate at each interim: placebo – 18.4%, active treatment – 16%.			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.93	0.96	0.99
Conditional power – future rates as protocol-specified	0.93	0.96	0.99
Conditional power – future rates a weighted average	0.93	0.96	0.99
PPS	0.67	0.84	0.98
(B) Assumed mortality rate at each interim: placebo – 18.4%, active treatment – 18.4%			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.02	0.01	0.00
Conditional power – future rates as protocol-specified	0.88	0.62	0.06
Conditional power – future rates a weighted average	0.81	0.27	0.00
PPS	0.27	0.08	0.00
(C) Assumed mortality rate at each interim: placebo – 16%, active treatment – 18.4%.			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.00	0.00	0.00
Conditional power – future rates as protocol-specified	0.80	0.15	0.00
Conditional power – future rates a weighted average	0.61	0.00	0.00
PPS	0.05	0.00	0.00

(i.e., at information fractions of 9%, 33%, and 67%). Conditional/predictive power is computed for every interim under each of three scenarios:

1. Observed event rates in the two treatment groups match the protocol-specified event rates.
2. Observed event rates in the two treatment groups are equal to the protocol-specified placebo event rate (no treatment difference).
3. Observed event rates are reversed–in the drotrecogin alfa (activated) group the protocol-specified placebo event rate is observed, and vice versa.

In computing the values in Table 1, we have assumed $\alpha = 0.0466$ as the significance level for the final analysis, an adjusted value specified in the design of ADDRESS, allowing for a nominal α-spend at each interim analysis so as to preserve an overall Type 1 error of 0.05.

Under the first scenario the performance of conditional power and PPS is similar, indicating support for continuing the trial, as expected. There is no difference between the three variations of conditional power since observed event rates are assumed to match the protocol-specified rates. Because PPS incorporates the uncertainty in future event rates that is inherent especially early in a trial, it tends to produce lower values than conditional power at early interim analyses, even under optimistic scenarios.

The second scenario assumes no observed effect of treatment, and a useful futility rule should suggest stopping the trial relatively early, but not before sufficient information has been collected to be reasonably sure that the desired treatment difference won't be seen by the end of the study. The PPS is relatively low at the first interim, but likely not low enough to warrant stopping at that point. The evidence to support a decision to stop is much greater, however, at the second interim and undisputable at the third. In contrast, conditional power under the assumption that future event rates are fixed at the observed rates places too much evidence on the observed data and might result in stopping a study too soon; quite the opposite is true for the two alternative ways of computing conditional power.

Under the third scenario, if safety considerations alone were not sufficient to lead to a decision to terminate the trial, the fact that observed event rates are reversed should clearly lead one to conclude futility as early as possible. Conditional power under the assumption that future event rates are fixed at the observed rates achieves this end, but neither of the two alternative approaches would yield the appropriate conclusion at the first interim. PPS suggests a low likelihood of success even by the first interim, and thus performs reliably without having to make additional explicit assumptions about future event rates as with conditional power.

The three cases above describe a situation in which a trial should not be stopped for futility, one in which the data would indicate ambiguity if the interim in question were early in the study but where the prudence of stopping becomes clearer as additional data is collected, and one in which very early stopping is warranted. Unlike PPS, no single version of conditional power addresses every one of these situations consistently well. PPS in essence performs like the most desirable formulation of conditional power under each scenario. At the same time we acknowledge the close connection between the two approaches and the fact that presentation of various contrasting formulations of conditional power may in

fact be more readily understood by some DMCs. Nonetheless, the properties and applications of PPS studied in this paper are germane topics for conditional power and, more generally, other means of assessing futility as well.

Following the recommendation of an external DMC appointed for this trial, ADDRESS was stopped in February 2004 for futility based on a low PPS of less than 5%. This recommendation was the result of an unplanned interim analysis of approximately 1500 patients, requested by the DMC after their first planned interim analysis. Observed 28-day mortality rates at the time were 17.3% and 19.7% in placebo and drotrecogin alfa (activated)-treated patients, respectively. Further discussion of the interpretation of the ADDRESS data is provided in Section 6.

Having examined scenarios for the ADDRESS trial in particular, we now explore characteristics of PPS in the context of large outcomes trials in general. We considered three hypothetical trials with a planned enrolment of 9000 patients each, for which interim analyses are planned after 1000, 3000, and 6000 patients have been enrolled. The trials were all designed to have 90% power to detect a statistically significant difference in event rates between treatment groups, but differ in their assumed placebo event rates: rates of 10, 25, and 50% were assumed for the placebo group, yielding assumed rates in the active treatment arm of 8.02, 22.1, and 46.6%, respectively, that produce differences detectable with 90% power in each case. These differences correspond to 19.8, 11.7, and 6.9% relative risk reductions, respectively. Table 2 shows the same comparisons as Table 1 for these three trials, with the exception that only scenarios where observed event rates at a given interim analysis are equal or the reverse of protocol-specified rates are considered. In computing the values in Table 2, a significance level of $\alpha = 0.045$ was assumed for the final analysis. The results are very similar to those in Table 1 and discussed in the previous section, indicating that similar behavior of PPS can be expected over a broad range of event rates.

So far we have focused on investigating PPS under various specific scenarios, but have not addressed the choice of appropriate thresholds which might be used to decide whether to continue or terminate a trial. To investigate this question we simulated 4000 data sets for each of the same three trial scenarios as discussed above, both under the null hypothesis of no treatment effect and under the alternative hypothesis of a 19.8, 11.7, and 6.9% relative risk reduction, respectively. We computed PPS at each interim for each simulated data set, and studied the resulting empirical distributions. Table 3 gives the 5th, 10th, 50th, and 95th percentiles of these distributions. Comparing PPS between null and alternative hypotheses, we note that the median values are well differentiated as expected, with probabilities tending to 0 under H_0 and to 1 under H_A. Nevertheless, there is considerable overlap in the distributions of plausible values of PPS when comparing H_0 and H_A. In order to avoid a Type II error by invoking a futility rule, one must choose thresholds that are sufficiently conservative. The power of a given study can facilitate the choice of such thresholds. Consider that a study with power $1 - \beta$ will find no significant difference between treatments $100\beta\%$ of the time, even if H_A is true. Since conclusions for a trial can only be based on the data observed, it is not only desirable to stop trials early in which there is no treatment effect in truth, but also those trials testing truly efficacious therapies, but unfortunately destined for a Type II error if left to continue. Considering therefore the PPS distributions generated under H_A with power 90%, it seems reasonable that the smallest 10% of

Table 2 PPS for simulated 9000-patient trials

Approach	Interim 1 $n = 1000$	Interim 2 $n = 3000$	Interim 3 $n = 6000$
(1A) Assumed event rate at each interim: placebo – 10%, active treatment – 10%			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.02	0.01	0.00
Conditional power – future rates as protocol-specified	0.83	0.58	0.05
Conditional power – future rates a weighted average	0.72	0.24	0.00
PPS	0.24	0.08	0.00
(1B) Assumed event rate at each interim: placebo – 8.02%, active treatment – 10%.			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.00	0.00	0.00
Conditional power – future rates as protocol-specified	0.72	0.13	0.00
Conditional power – future rates a weighted average	0.45	0.00	0.00
PPS	0.03	0.00	0.00
(2A) Assumed event rate at each interim: placebo – 25%, active treatment – 25%			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.02	0.01	0.00
Conditional power – future rates as protocol-specified	0.83	0.57	0.05
Conditional power – future rates a weighted average	0.73	0.24	0.00
PPS	0.24	0.08	0.00
(2B) Assumed event rate at each interim: placebo – 22.1%, active treatment – 25%			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.00	0.00	0.00
Conditional power – future rates as protocol-specified	0.71	0.13	0.00
Conditional power – future rates a weighted average	0.45	0.00	0.00
PPS	0.03	0.00	0.00
(3A) Assumed event rate at each interim: placebo – 50%, active treatment – 50%			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.02	0.01	0.00
Conditional power – future rates as protocol-specified	0.83	0.58	0.05
Conditional power – future rates a weighted average	0.73	0.24	0.00
PPS	0.24	0.08	0.00
(3B) Assumed event rate at each interim: placebo – 46.6%, active treatment – 50%			
Conditional power – future rates fixed at x_T/n_T, x_C/n_C	0.00	0.00	0.00
Conditional power – future rates as protocol-specified	0.71	0.13	0.00
Conditional power – future rates a weighted average	0.46	0.00	0.00
PPS	0.03	0.00	0.00

Table 3 Distribution of PPS (Predictive Probability of Success) under H_0 and H_A for simulated 9000-patient trials

		Interim 1 ($n = 1000$) Percentiles				Interim 2 ($n = 3000$) Percentiles				Interim 3 ($n = 6000$) Percentiles			
		5th	10th	50th	95th	5th	10th	50th	95th	5th	10th	50th	95th
Trial 1	H_0	0.01	0.02	0.24	0.87	0.00	0.00	0.08	0.76	0.00	0.00	0.00	0.56
	H_A	0.10	0.18	0.68	0.98	0.16	0.27	0.83	1.00	0.19	0.39	0.97	1.00
Trial 2	H_0	0.01	0.02	0.24	0.86	0.00	0.00	0.08	0.72	0.00	0.00	0.00	0.50
	H_A	0.11	0.20	0.66	0.99	0.14	0.28	0.81	1.00	0.17	0.36	0.96	1.00
Trial 3	H_0	0.01	0.02	0.25	0.86	0.00	0.00	0.08	0.73	0.00	0.00	0.00	0.51
	H_A	0.11	0.19	0.67	0.99	0.12	0.27	0.82	1.00	0.15	0.36	0.97	1.00

Values in each table cell correspond to 5th, 10th, 50th and 95th percentiles of the empirical distribution of PPS values observed over 4000 simulations.

For Trial 1, H_0: placebo event rate (π_C) = active treatment event rate (π_T) = 0.10
H_A: $\pi_C = 0.10$, $\pi_T = 0.0802$;

For Trial 2, $H_0 : \pi_C = \pi_T = 0.25$ $H_A : \pi_C = 0.25$, $\pi_T = 0.221$;

For Trial 3, $H_0 : \pi_C = \pi_T = 0.50$ $H_A : \pi_C = 0.50$, $\pi_T = 0.466$.

PPS values should typically reflect those values that, if observed, would be indicative of a trial bound for a Type II error; larger PPS values would likely not support a decision to stop for futility. Examining the 10th percentiles of these distributions in Table 3 hence suggests that appropriate thresholds for stopping for futility are in the range of 20–40%, depending on the observed information fraction at the interim in question. Of course one might choose a more conservative threshold in order to adjust for multiplicity; there is a straightforward parallel between significance level adjustments to control Type I error in multiple interim analyses for efficacy, and similar adjustments to control Type II error in multiple tests for futility.

In generating the PPS distributions for the second and third interim analyses, we included all simulated trials in our calculations rather than excluding those that might have been terminated for reasons of superior efficacy at an earlier interim. Such exclusions would necessarily have depended on the choice of a specific efficacy stopping rule, detracting from generalizability of the results. Therefore, the median and larger quantiles of our empirical PPS distributions may somewhat overestimate the true values of these parameters. Note, however, that in determining thresholds for decision rules, primarily the left tail of the distribution of PPS values is of interest, that is, the part of the distribution least affected by using all simulated trials in the computations. Moreover, in clinical practice the numerous risks inherent in terminating trials prematurely in order to claim efficacy are gaining increased appreciation (Montori et al., 2005; Sashegyi, 2004), such that current trends in data monitoring appear to be favoring either very stringent efficacy stopping rules or precluding the possibility of stopping for efficacy altogether in favor of exclusive focus on safety and futility.

5. A NOTE ON OVERDISPERSION

The approaches described in Section 3 rely on the assumption that conditional on a rate parameter π, the number of events X observed in n patients in a particular treatment group is binomially distributed, that is,

$$X \mid \pi \sim \text{Bin}(n, \pi).$$

Now in the setting of large outcomes trials, X is bound to be the sum of event counts from a number of clusters or regions, such as various clinical sites at the least or even entire countries, as in a multinational trial. Thus,

$$X = X_1 + \cdots + X_K \quad \text{and} \quad X_i = X_{i1} + \cdots + X_{in_i},$$

where X_i is the number of events observed in cluster i, $i = 1, \ldots, K$, and $X_{ij} \sim \text{Bin}(1, \pi)$, $j = 1, \ldots, n_i$, is a binary indicator of whether patient j in cluster (e.g., country) i has had an event. For the purpose of illustration, consider a global trial like ADDRESS or RUTH; these studies were conducted in 34 and 26 countries, respectively. Given the likely heterogeneity in event rates from country to country, it seems reasonable to expect two patients from the same country to be more similar with respect to their likelihood of having an event than two patients from different countries. Thus, X_{ij} and $X_{ij'}$ are positively correlated, leading to an overdispersed binomial distribution with larger-than-nominal variance. Specifically, assuming X_i and $X_{i'}$ to be independent but $\text{Corr}(X_{ij}, X_{ij'}) = \rho$, $0 < \rho < 1$, $j \neq j'$, it follows that the binomial mean is unaffected since $\text{E}(X \mid \pi) = n\pi$, but the variance is inflated in that

$$\sigma^2 = \text{Var}(X \mid \pi) = n\pi(1-\pi) + \rho\pi(1-\pi) \sum_{i \in \{1,\ldots,K\}} n_i(n_i - 1),$$

where $n = n_1 + \cdots + n_K$. If $n_1 = \cdots = n_K = m$, $\text{Var}(X \mid \pi)$ simplifies to $n\pi(1-\pi)\{1 + \rho(m-1)\}$, with $1+\rho(m-1)$ emerging as the familiar extra-binomial variance inflation factor for cluster-correlated data with clusters of size m; ρ is often referred to as the intra-class or intra-cluster correlation. Hence prior to proceeding with any calculations that assume $X \mid \pi \sim \text{Bin}(n, \pi)$, it may be of interest to test H_0: $\rho = 0$, or equivalently, the hypothesis

$$H_0 : T = \sigma^2/(n\pi(1-\pi)) = 1. \tag{7}$$

An empirical estimate of the observed variance $\sigma^2 = \text{Var}(X \mid \pi)$ is given by

$$s^2 = \sum_{i \in \{1,\ldots,K\}} (x_i - n_i p)^2,$$

where $p = x/n$, with x_i and x representing realized values of X_i and X. Further, replacing π with p in Equation (7), a resampling technique such as the bootstrap (Davison and Hinkley, 1997) can be easily applied to obtain the empirical distribution of the estimate of T, and thus a test of H_0: $T = 1$. At interim time points, in both ADDRESS and RUTH T was estimated to be significantly larger than 1, indicating an observed variability 2.5 to 3 times larger than nominal

binomial variation. Indeed, country-to-country variation in event rates in global outcomes trials should be expected to be the norm rather than the exception. More generally, one should anticipate overdisperion in any setting in which data are clustered. Extraneous variation even between clinical sites can be significant. And certainly cluster-randomized trials in which treatments are assigned not to individual experimental units but to groups of subjects, such as families, schools, even entire communities, imply by their very design that overdispersion is likely to be a factor and that within-cluster correlation will need to be taken into account.

The question of interest with regard to assessing futility by means of computing PPS as described above (and with regard to power calculations in general) concerns the impact of overdispersion, and when it really matters. If the goal of a study is solely to estimate a treatment effect, overdispersion should not impact this assessment in trials randomized within country (cluster), since such randomization allows for within-cluster treatment comparisons that are not confounded with cluster-to-cluster variability. To demonstrate that this is true, carrying on with the development above, assume the correlation structure described and that patients are randomized to two treatment groups T or C within country. Thus, $\mathrm{Corr}(X_{Tij}, X_{Cij'}) = \rho, 0 < \rho < 1$, for any j, j'. In computing power or PPS $\mathrm{E}(p_T - p_C)$ and $\mathrm{Var}(p_T - p_C)$ are the two factors of key importance. It is easy to show that $\mathrm{E}(p_T - p_C)$ is unaffected by overdispersion, and assuming for ease of illustration the simple case in which $n_1 = \cdots = n_K = m$, with n patients per treatment arm, it follows that

$$\mathrm{Var}(p_T - p_C) = (2/n)\pi(1 - \pi)$$

under nominal binomial variation ($\rho = 0$), and

$$\mathrm{Var}(p_T - p_C) = (2/n)\pi(1 - \pi)(1 - \rho)$$

if the data are overdispersed ($\rho > 0$). Considering the inflation factor $1 + \rho(m - 1)$, one can appreciate that even small values of ρ can lead to significant variance inflation in $\mathrm{Var}(X \mid \pi)$. Thus in a global trial involving several thousand patients in total and an average of, say, 200 patients per country, one should expect on the one hand an approximately three-fold variance inflation in terms of $\mathrm{Var}(X \mid \pi)$ even with $\rho = 0.01$. On the other hand, this would have an imperceptible impact on $\mathrm{Var}(p_T - p_C)$.

To investigate this impact numerically, we generated data under the three simulated trial scenarios presented in Section 4, keeping all assumptions the same but introducing extreme overdispersion into the data. Operationally, we assumed in every scenario that the 4500 patients in each treatment arm represented 20 countries contributing 225 patients each to the study. Correlation between two patients in the same country was achieved by postulating a normally-distributed country-level random effect impacting event rates on the logit scale. Assuming a random effects variance of zero corresponds to nominal dispersion in this case. We show comparative results for simulated data assuming a random effects variance of 0.5, which implies extremely overdispersed data with a variance inflation factor T ranging from 5 to 14, depending on the assumed placebo event rates. Table 4 displays these results for the second and third interim analyses of the three simulated

Table 4 Distribution of PPS under H_0 and H_A for simulated 9000-patient trials–overdispersed data

		Interim 2 ($n = 3000$) Percentiles				Interim 3 ($n = 6000$) Percentiles			
		5th	10th	50th	95th	5th	10th	50th	95th
Trial 1	H_0	0.00	0.00	0.08	0.71	0.00	0.00	0.00	0.51
	H_A	0.16	0.29	0.84	1.00	0.18	0.40	0.97	1.00
Trial 2	H_0	0.00	0.00	0.08	0.74	0.00	0.00	0.00	0.49
	H_A	0.13	0.25	0.81	1.00	0.15	0.33	0.96	1.00
Trial 3	H_0	0.00	0.00	0.08	0.71	0.00	0.00	0.00	0.47
	H_A	0.12	0.23	0.80	1.00	0.12	0.30	0.95	1.00

Comparative results for the same trial scenarios as in Table 3, simulated using a logistic-normal random effects model producing highly overdispersed data. Variance inflation factors T range from approximately 5 (Trial 3) to 14 (Trial 1), depending on the assumed event rates.

trials. Barring sampling error there is very little difference in the quantiles shown, as compared to the corresponding values in Table 3.

Fortunately, the estimation of treatment effect is the primary objective of most clinical trials, and randomization is typically carried out at the level of the experimental unit (subject), *within*-cluster, as described above. Hence no adjustment is necessary to the computation of PPS in most settings. On the other hand, if interest lies in prediction of an event rate in a single arm of a study in which a comparison is made to, say, a historical control rate, overdispersion must be taken into account. The same applies to all cluster-randomized trials, as indicated above. Techniques such as stratification to help ensure balance between treatment groups generally will not address the problem of overdispersion because these measures are typically implemented within each cluster and do not deal with extraneous cluster-to-cluster variability. Even in cluster-randomized trials, stratification or dynamic randomization can virtually guarantee balanced treatment groups with respect to important covariates, but will not address the issue of variability within a treatment arm. If the assumption that $X \mid \pi \sim \text{Bin}(n, \pi)$ no longer holds because X is overdispersed, one cannot assume that $\pi \mid X = x \sim \text{Beta}(a + x, b + n - x)$.

Heuristically, in computing PPS in the cluster-randomized setting the distributions of event rates governing future observations should have variances similarly inflated as those of the data giving rise to these distributions. Interestingly, in the beta-binomial framework, if x and n are large compared to a and b,

$$\begin{aligned}\text{Var}(\pi \mid x) &= (1 + a + b + n)^{-1}((a + x)/(a + b + n))((b + n - x)/(a + b + n)) \\ &\approx p(1 - p)/n, \quad p = x/n \\ &= n^{-2}\text{Var}(X \mid p).\end{aligned}$$

This would suggest that insofar as the variance of $\pi \mid x$ is a function of the variance of $X \mid p$, observations drawn from this posterior distribution (as in Steps 1 and 3 of Sections 3.1 and 3.2, respectively) could be appropriately adjusted in the presence of overdispersion using the same variation inflation factor that can be calculated

for the data X via (7). Again, avoiding subscripts for the sake of clarity, one naïve adjustment to the sampled values π^* obtained in Steps 1 and 3 of Sections 3.1 and 3.2 would be to replace π^* with

$$\pi_o^* = E(\pi \mid x) + \sqrt{T}(\pi^* - E(\pi \mid x)),$$

estimating $E(\pi \mid X)$ and T from the observed data. Clearly, accounting for overdispersion in this manner in computing PPS becomes problematic if the π^* are close to 0 or 1, or if T is large, since π_o^* may then well fall outside the range (0, 1). Moreover, we have not compared the properties of PPS using this or similar heuristic adjustments with overdispersed data to those in the setting of nominally dispersed data. Suffice it to say that just as overdispersion should be anticipated at the design stage of trials and, if need be, accounted for explicitly in power calculations, so too should the computation of PPS at interim time points be adjusted in a sensible manner, when indicated.

6. DISCUSSION

We have presented a Bayesian approach for facilitating the construction of futility rules based on predictive probability of success, which is especially useful for large endpoint-based clinical trials in which computation of individual probabilities for particular future outcomes is not feasible. This approach assigns means and variances of the distributions of event rates governing future outcomes to be intuitive functions of the data, which yields reasonable results under a broad range of conditions.

We have given some indication of appropriate thresholds for PPS, which might be used to determine whether or not to terminate a trial due to futility. The most reasonable choice of threshold in any given situation will depend on the power of the study and the relative seriousness of Type II error inflation-increasing the threshold will increase the chances of stopping a futile study, but also result in loss of power. Another consideration in choosing specific futility thresholds is the number and timing of interim analyses at which futility assessments are made. Based on Table 3, small quantiles of the distribution of PPS vary only moderately with respect to the information fraction at a given interim analysis, but certainly larger quantiles will shift more quickly toward 0 under the null and 1 under the alternative hypothesis. Thus, for large pivotal registration trials in which loss of power is highly undesirable, a futility threshold in the order of 20% would seem reasonable, whereas for smaller studies higher values might be acceptable. In any case, futility assessments are clearly riskier if undertaken early in a trial. Varying PPS thresholds for different interim analyses, that is, gradually choosing less conservative values with increasing information fractions is a reasonable strategy. Moreover, the simulation of relevant trial scenarios and construction of empirical distributions of PPS to help guide the optimal choice of thresholds is not difficult, and is recommended in any case as part of the planning process of a new trial that is to involve a PPS-based futility rule.

The appealing flexibility that PPS offers over conditional power is reflected in the fact that one need not make fixed assumptions about event rates governing

future observations. However, the approach still has the limitation of assuming that the *process* governing the occurrence of events is constant over time, that is, that mean event rates in each treatment group do not change over time. This may be unreasonable in some cases. In device trials and to a lesser extent with certain therapeutic interventions, for example, it is possible for sites participating in a study to undergo a learning period, during which the benefit of a new procedure relative to a standard may be obscured until such time as the new intervention is consistently and optimally applied by site personnel (Halm et al., 2002; Katzan et al., 2000). In such cases, subgroup analysis according to the sequence of enrolment of each patient at his or her site (e.g., the first two patients at each site versus the rest of the patients) might reveal no efficacy in patients enrolled early, offset by a significant effect in the remaining patients. A closer analysis revealed this very phenomenon in the ADDRESS data. At the time the DMC made its recommendation to stop the trial, a total of approximately 2600 patients had already been enrolled in the trial. The final mortality rates were 17.0% and 18.5% in placebo and drotrecogin alfa (activated)-treated patients, respectively. The observed 1.5% mortality difference in disfavor of drotrecogin alfa (activated) appeared to be explained by markedly higher mortality confined to the first patient enrolled at each site. Upon removing these first patients, mortality rates were nearly identical in the two treatment arms, with the removal of the first 2, 3, and 4 patients at each site leading to progressively larger mortality differences in favor of drotrecogin alfa (activated). This apparent effect reversal was not explained by any observed imbalances in patient characteristics. Had such a sequence effect been anticipated in ADDRESS, the futility rule could have been appropriately adjusted, for example, by omitting patients enrolled early in the trial or at a site in conditioning on observed data. The assumption of homogeneity in key parameters over time could also be tested explicitly before evaluating PPS as described above.

In a clinical research environment in which the role of patient safety, ethics, and also cost are ever increasing in importance, rigorous interim monitoring of large trials will become perhaps the most critical element of the trial conduct process. Appropriate efficacy and safety stopping rules alone may not help a DMC make a potentially important decision on the utility of continuing a study when there is neither an efficacy signal nor a sufficient safety concern in the data to warrant discontinuation. We have described an approach to assessing futility with desirable properties that can help close this gap and complete the scope of decision guidelines with which one should approach interim analyses. As a final note, all futility criteria should be viewed as delineating conditions under which a recommendation to terminate a trial should be seriously considered; nevertheless, without detracting from their usefulness, these criteria alone may not necessarily be sufficient to warrant that recommendation in fact. Other trial aspects and the observed data in its entirety should always be taken into account in forming a final recommendation.

ACKNOWLEDGMENTS

The authors are grateful for the comments of a referee, which helped to improve the quality and clarity of this manuscript.

REFERENCES

Abraham, E., Laterre, P. F., Garg, R., Levy, H., Talwar, D., Trzaskoma, B., Francois, B., Guy, J. S., Brueckmann, M., Rea-Neto, A., Rossaint, R., Perrotin, D., Sablotzki, A., Arkins, N., Utterback, B., Macias, W for the ADDRESS Study Group. (2005). Drotrecogin alfa (activated) in adult severe sepsis patients at low risk of death. *New England Journal of Medicine* 353:1332–1341.

Barrett-Connor, E., Mosca, L., Collins, P., Geiger, M. J., Grady, D., Kornitzer, M., McNabb, M., Wenger, N. for the RUTH Trial Investigators. (2006). Effects of raloxifene on cardiovascular events and breast cancer in postmenopausal women. *New England Journal of Medicine* 355:125–137.

Bernard, G. R., Vincent, J. L., Laterre, P. F., LaRosa, S. P., Dhainaut, J. F., Lopez-Rodriguez, A., Steingrub, J. S., Garber, G. E., Helterbrand, J. D., Ely, E. W., Fisher, C. J. (2001). Recombinant Human Protein C Worldwide Evaluation in Severe Sepsis (PROWESS) Study Group. Efficacy and safety of recombinant human activated protein C for severe sepsis. *New England Journal of Medicine* 344:699–709.

Casella, G., George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician* 46:167–174.

Choi, S. C., Smith, P. J., Becker, D. P. (1985). Early decision in clinical trials when the treatment differences are small. *Control Clin. Trials* 6:280–288.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Freidlin, B., Korn, E. L. (2002). A comment on futility monitoring. *Control Clin. Trials* 23:355–36.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (1995). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall.

Halm, E. A., Lee, C., Chassin, M. R. (2002). Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. *Annals of Internal Medicine* 137:511–520.

Halperin, M., Lan, K. K. G., Ware, J. H., Johnson, N. J., DeMets, D. L. (1982). An aid to monitoring in long-term clinical trials. *Control Clin. Trials* 3:311–323.

Herson, J. (1979). Predictive probability early termination plans for phase II clinical trials. *Biometrics* 35:775–783.

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., Vittinghoff, E., for the HERS Research Group. (1988). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association* 280:605–613.

Jennison, C. J., Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman & Hall/CRC.

Johns, D., Andersen, J. S. (1999). Use of predictive probabilities in phase II and phase III clinical trials. *J. Biopharm. Stat.* 9(1):67–79.

Katzan, I. L., Furlan, A. J., Lloyd, L. E., Frank, J. I., Harper, D. L., Hinchey, J. A., Hammel, J. P., Qu, A., Sila, C. A. (2000). Use of tissue-type plasminogen activator for acute ischemic stroke. *Journal of the American Medical Association* 283:1151–1158.

Montori, V. M., Devereaux, P. J., Adhikari N. K. J. et al. (2005). Randomized trials stopped early for benefit. *Journal of the American Medical Association* 294(17):2203–2209.

Sashegyi, A. (2004). Too much success, much too soon? *Good Clinical Practice Journal* 11(10):8–9.

Spiegelhalter, D. J., Freedman, L. S., Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power. *Controlled Clinical Trials* 7:8–17.

Van der Tweel, I., van Noord, P. A. H. (2003). Early stopping in clinical trials and epidemiologic studies for "futility": Conditional power versus sequential analysis. *Journal of Clinical Epidemiology* 56:610–617.

Wittes, J. (2000). Data safety monitoring boards: a brief introduction. *Biopharmaceutical Report* 8(1):1–7.

Women's Health Initiative Investigators' Writing Group. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *Journal of the American Medical Association* 288:321–333.