A Method for Predicting a Binomial Response Rate in a Bayesian Interim Analysis of a Single-Arm Trial when Observing Responses and Failures Requires Prolonged Follow-up

Gary Hantsbarger
Astellas Pharma Global Development, 1 Astellas Parkway, Northbrook, IL 60069

Abstract
Initial investigation of the efficacy of a drug may involve a single-arm study in which the response rate of subjects receiving the drug is compared to a pre-defined minimum rate, below which the drug is deemed not worth further study. During recruitment, interim futility analyses may be used to determine whether the probability of a successful trial is high enough to justify continuing the study. When response or failure is quickly determined, Bayesian interim analyses for such studies typically use a beta distribution as a prior for the response rate. However, if observing a response requires prolonged follow-up, this model is not adequate. A Bayesian interim analysis method which takes into account both time to failure and time to response, while respecting the binomial nature of the final analysis, will be described. The method is based on decomposing a beta prior for a binomial response/failure probability into a Dirichlet distribution describing probabilities of failure and response among multiple time intervals.

Key Words: Bayesian, futility analysis, Dirichlet distribution, predicted probability of success

1. Introduction

Initial investigations of the efficacy of a new cancer treatment often involve single arm studies in which the response rate of subjects receiving the new therapy is compared to a pre-defined minimum response rate. The minimum response rate is defined such that treatments which cannot be shown to produce response rates at least as high as the minimum can be considered as not worth pursuing in larger trials. Single arm trials can be justified when other treatments available for the study's target population are only minimally beneficial, or have especially onerous side effects.

If the number of subjects to be recruited is large or the time required for recruitment is long, both ethical and financial reasons motivate the use of interim futility analyses to determine, prior to the completion of recruitment, whether the probability of a successful trial is high enough to justify continuing the study.

It is assumed that subjects have an unknown true response rate of π. The number of responses seen in a trial, given the response rate π, is assumed to have a binomial distribution. That is,

$$f_r(r \mid \pi) = \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

with $f_r(r|\pi)$ being the probability of observing exactly r responses out of n patients treated, given π, the unknown true response rate for the subject population. In a Bayesian context, the conjugate prior distribution for the response rate π is the beta distribution, with the probability distribution function

$$f_\pi(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}.$$

with parameters $\alpha > 0$ and $\beta > 0$. The mean of this distribution is $\alpha/(\alpha + \beta)$. The magnitude of $\alpha + \beta$ relate to the strength of prior belief about the event rate. Typically, for a new treatment's first clinical trial, these parameters are set such that $\alpha + \beta \leq 2$.

After n_1 patients have been treated and r_1 responses have been observed the posterior distribution of π is

$$f_\pi(\pi \mid r_1, n_1) = \frac{\Gamma(\alpha+\beta+n_1)}{\Gamma(\alpha+r_1)\Gamma(\beta+n_1-r_1)}\pi^{\alpha+r_1-1}(1-\pi)^{\beta+n_1-r_1-1}$$

If outcomes for n_2 patients are yet to be observed, then the distribution of total responses which will be observed is expressed as a beta-binomial distribution, with the distribution function

$$f_\pi(r_2 \mid \pi, r_1, n_1) = \frac{\Gamma(\alpha+\beta+n_1+n_2)}{\Gamma(\alpha+r_1+r_2)\Gamma(\beta+n_1+n_2-r_1-r_2)}\pi^{\alpha+r_1+r_1-1}(1-\pi)^{\beta+n_1+n_2-r_1-r_2-1}$$

From this, the probability that enough total responses will be seen by the end of the study to declare the study a success (e.g., by rejecting a null hypothesis that $\pi \leq \pi_{null}$) can be estimated. Thus, we can recruit, treat, and examine n_1 patients and then determine whether there is enough probability of eventual success to make the recruitment of a further n_2 patients worthwhile. This is the essence of Bayesian interim analysis for Phase II studies as developed by (Thall and Simon, 1994).

This assumes that the outcomes of all of the first group of n_1 patients, whether responses or failures, are observed before any of the any of the second group of n_2 patients are treated. If each patient's outcome is known soon after the start of treatment then it may be feasible to pause recruitment after the first n_1 patients, determine the outcome for each, and then, if warranted, restart recruitment.

However, there may be an extended period between the start of treatment for a patient and determination of the patient's treatment outcome. Pausing recruitment may be impractical. If recruitment cannot be paused, and the rate of recruitment is not very slow, then many of the subjects comprising the second group may be already recruited before the all outcomes in the first group are known. It will be of little benefit to either patients or to the sponsor if the second group is already recruited when the analysis of the first group has no chance of showing efficacy.

An alternative is to perform an interim analysis when the outcomes of n_1 patients have been determined, regardless of whether these subjects are the first n_1 patients recruited. However, these outcomes may not be representative of all outcomes which would be seen if the trial continued to completion. Identifying a patient as a responder may require blood tests, biopsies, or scans which can only be done a limited number of times during follow-up and which then have to be confirmed by repeated procedures, while patients may be identified as treatment failures due symptomatic progression or intolerable side effects which can be become immediately evident at any time during follow-up.

Also, if the study population consists of patients with tumors susceptible to the experimental treatment and others with tumors not susceptible, failures may occur quite quickly, whereas responses may follow a long period of stable disease before becoming evident.

In either case, potentially useful information from subjects who have been recruited, but not yet evidenced an outcome would be lost to the interim analysis. Therefore, an interim analysis which takes into account both time to failure and time to response, while at the same time respecting the binomial nature of the final analysis, would be useful.

This paper presents a method for monitoring a single arm trial when there may be a long and variable delay between the beginning of treatment and the determination of response or failure, and when the distribution of times to treatment failure may differ from times to response. It assumes that follow-up times can be divided into a limited number of defined intervals, as would be the case when treatment is given in regular cycles. Development of the method is based on decomposing the beta prior for a binomial response/failure probability into a Dirichlet distribution describing probabilities of response and failure among multiple time intervals.

The following section explores the potential severity of the bias of continuing to assume a binomial distribution of responses in this situation. Section 3 develops the method, and Section 4 provides an example based on hypothetical data.

2. Bias of the Naïve Method

Let us suppose that a trial is planned with n_{total} as the total sample size and that A subjects must experience a response to the experimental treatment for the study to be considered a success. Suppose also that after n_1 subjects are known to be either responders or nonresponders an interim analysis will be performed and the study will be stopped for futility if the predictive probability of success $p_{success}$ is less than a futility limit p_{futile}.

The study is planned with the assumption that response has a binomial distribution with parameter π, which in turn has a prior distribution Beta(α_o, β_o).

As an example, let n_{total} = 140 subjects, n_1 = 30 subjects, A = 22 subjects, and p_{futile} = 0.1. If we assume for π a relatively non-informative prior distribution of Beta(0.2, 1.8), then the study will be stopped at the interim analysis if there are fewer than three successes out of the first 30 subjects, if the stopping decision is based on the assumption that the relative proportion of failures and successes is constant. If there is no time delay or bias in the reporting of outcomes, then the number of successes at the interim analysis and the number of successes after the interim analysis are independent binomial variables. On this basis, the probabilities of a positive outcome and early stopping can be computed as a function of the true (unobserved) proportion of responders in the population.

These probabilities will hold regardless of the rate of recruitment or the time from start of study to the outcome, as long as the distributions of times to outcome for responding subjects is the same as that for non-responding subjects. But this will not be the case if these distributions differ.

For example, suppose that the interim analysis will be done on the first n_1 subjects with known outcomes, and that for the 100π% of the subjects who respond to treatment a minimum of c cycles must pass before a response can be seen. Let t_{cycle} be the time it takes for one cycle of treatment to be completed. And suppose that subjects enter the study at a constant rate of r subjects per t_{cycle}. If $t_{interim}$ is the time from the beginning of the trial to the interim analysis and $t_{interim}/t_{cycle} < c$, then no subjects will have had time to show a response by the time of the interim analysis and the study will be stopped for failure regardless of the true value of π. Furthermore, approximately $(t_{interim}/t_{cycle})\cdot r$ will have been recruited to a study which may have been incorrectly terminated.

3. Restructuring the Binomial Model as a Multinomial Model

A Dirichlet distribution is a generalization of a beta distribution to multivariate parameters. It describes a distribution of a vector of nonnegative random variables which sum to one. Whereas the beta distribution serves as the conjugate prior for a binomial distribution, the Dirichlet distribution serves as conjugate prior for a multinomial distribution. For a vector of size n, there are n positive parameters, which may be written as $(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$. The Dirichlet distribution has the following properties which will prove useful. If $(\theta_1, \theta_2, \theta_3, \ldots, \theta_n) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$ then the sum of any subset of the θ_i's will have a beta distribution. For example, $\theta_1 + \theta_2 \sim \mathrm{Beta}(\alpha_1 + \alpha_2, \sum_{i=3}^{n} \alpha_i)$. Furthermore, if any subset of the vector of variables is discarded, the remaining variables, if normalized by their sum, will still have a Dirichlet distribution which is independent of the discarded variables. For example, if we discard θ_1 from the vector then

$$\left(\frac{\theta_2}{1-\theta_1}, \frac{\theta_3}{1-\theta_1}, \ldots, \frac{\theta_n}{1-\theta_1}\right) \sim \mathrm{Dir}(\alpha_2, \alpha_3, \ldots, \alpha_n)$$

and

$$\theta_1 \perp\!\!\!\perp \left(\frac{\theta_2}{1-\theta_1}, \frac{\theta_3}{1-\theta_1}, ,\ldots, \frac{\theta_n}{1-\theta_1}\right).$$

Now suppose that there are T intervals within the study, perhaps corresponding to T cycles of treatment, and let the probability of failure in interval t be $\theta_{F,t}$, and the probability of a response during interval t be $\theta_{R,t}$. (Note that these are marginal probabilities. They are *not* the conditional probabilities of failure or response in interval t for subjects who have already completed t-1 intervals.) Let the vector of these probabilities,

$$(\theta_{R,1}, \theta_{R,2}, \theta_{R,3}, \ldots, \theta_{R,T}, \theta_{F,1}, \theta_{F,2}, \theta_{F,3}, \ldots, \theta_{F,T}),$$

have a Dirichlet distribution with parameters

$$(\alpha_{R,1}, \alpha_{R,2}, \alpha_{R,3}, \ldots, \alpha_{R,T}, \alpha_{F,1}, \alpha_{F,2}, \alpha_{F,3}, \ldots, \alpha_{F,T}).$$

It is possible that observing a response during interval t may depend upon an assessment which is not performed during interval t. In this case, $\alpha_{R,t} = 0$ by definition. The same could be said for the probability of failure during an interval.

If the total response rate for the trial is modeled with a prior distribution Beta(α, β), then any model in which

$$\alpha = \sum_t \alpha_{R,t}$$

and

$$\beta = \sum_t \alpha_{F,t}$$

is consistent with the overall response model.

The choice of the Dirichlet parameters, like the choice of beta distribution parameters, depends upon how confident one is in the new treatment, or how cautious one wishes to be in stopping or continuing the trial at interim analyses. But additionally one may include in the prior distribution expectations about the time of response or failure. For example, if one expects that the effects of the treatment may require more than one cycle to manifest, while side effects will tend to eliminate patients during the first treatment cycle then one might chose a relatively high value for $\alpha_{F,1}$, and a relatively low value for $\alpha_{R,1}$.

A neutral choice might be to let $\alpha_{R,t} = \alpha S^{-1}$ and $\alpha_{F,t} = \beta T^{-1}$, where S and T are the number of intervals during which there are assessments which might produce a response or failure, respectively. Another neutral choice might be to initially assume a constant hazard for response or failure across intervals up until a final point T at which any subjects still in the study would be regarded as treatment failures. In this case we would use for the parameters of the initial prior

$$\alpha_{R,t} = \alpha_{R,1}(1-\alpha_{R,1}-\alpha_{F,1})^{t-1}$$

and

$$\alpha_{F,1} = \alpha_{F,1}(1-\alpha_{R,1}-\alpha_{F,1})^{t-1}$$

for $t < T$ where $\alpha_{F,1}$, $\alpha_{R,1}$, and $\alpha_{F,T}$ can be any solution to the equations

$$\alpha = \sum\nolimits_{t=1}^{T-1} \alpha_{R,1}(1-\alpha_{F,1}-\alpha_{R,1})^{t-1}$$

and

$$\beta = \sum\nolimits_{t=1}^{T-1} \alpha_{F,1}(1-\alpha_{F,1}-\alpha_{R,1})^{t-1} + \alpha_{F,T} .$$

If $\theta_{F,t}$ and $\theta_{R,t}$ are the marginal probabilities for failure and response in interval t, hazards can be defined for failure and response by

$$\phi_{F,k} = \theta_{F,k} \Big/ \left(1 - \sum_{i=1}^{k-1}(\theta_{F,i}+\theta_{R,i})\right)$$

and

$$\phi_{R,k} = \theta_{R,k} \Big/ \left(1 - \theta_{F,k} - \sum_{i=1}^{k-1}(\theta_{F,i}+\theta_{R,i})\right) .$$

Then $\phi_{F,k}$ has the prior distribution Beta($\alpha_{F,k}$, $\alpha_{R,k} + \Sigma_{j>k}(\alpha_{R,j} + \alpha_{F,j})$), while $\phi_{R,k}$ has the prior distribution Beta($\alpha_{R,k}$, $\Sigma_{j>k}(\alpha_{R,j} + \alpha_{F,j})$). Furthermore, all of the phi's are independent, and the probability of a response occurring in interval k can be expressed as

$$\theta_{R,k} = \prod_{i=1}^{k}(1-\phi_{F,i})\prod_{i=1}^{k-1}(1-\phi_{R,i})\phi_{R,k} .$$

If we have observed n_i subjects at the ith interval, with f_i failures and r_i responses, then the posterior distribution on $\phi_{F,i}$ is Beta($\alpha_{F,i} + f_i$, $\alpha_{R,i} + \Sigma_{j>i}(\alpha_{R,j} + \alpha_{F,j}) + n_i - f_i$) and the posterior distribution of $\phi_{R,i}$ is Beta($\alpha_{R,i} + r_i$, $\Sigma_{j>i}(\alpha_{R,j} + \alpha_{F,j}) + n_i - f_i - r_i$).

Thus, based on observed data we can get posterior distributions for the probabilities of having a response or a failure at each cycle, conditional on not having had a previous response or failure.

Once posterior distributions for the phi's have been determined, the overall response rate can be expressed as a function of independent variables which all have known posterior distributions. If ORR is the overall response rate then

$$\text{ORR} = \sum_{k=1}^{n}\left(\prod_{i=1}^{k}(1-\phi_{F,i})\prod_{i=1}^{k-1}(1-\phi_{R,i})\phi_{R,k}\right).$$

We can draw from the posterior distributions for the phi's to get an overall probability of a response, and then draw from a binomial distribution using that probability to get an overall number of responses. If this is repeated a large number of times, the result is an estimate of the posterior distribution of the overall number of responses. This can be used to estimate the predictive probability of a successful trial.

If complete information on response or failure were known for subjects included in the interim analysis, then the predictive probability determined from the above analysis would be the same as the predictive probability resulting from an interim analysis based on the binomial distribution with a beta prior, because in this case the two methods are algebraically equivalent.

4 A Hypothetical Example

Consider a single-arm phase 2 cancer trial. Two hundred subjects will be given an experimental treatment. There are 5 treatment periods, each with multiple treatment cycles. At the end of each period, a subject can be classified into one of three possible outcomes:

- Response (success)
- Disease progression (failure)
- Stable disease.

Subjects with stable disease continue to the next treatment period, while subjects experiencing disease progression are withdrawn from the study. Based on the opinions of key experts and regulatory officials, it is determined that the treatment will be worth pursuing in larger-scale trials if the observed response rate is at least 26%. Hence, success for the study is defined as 52 or more responders out of 200 subjects. Also, Beta(0.2, 1.8) is selected as a reasonable prior distribution for the binomial response rate.

After a set period of time, a first futility analysis is to be carried out. At this time, 33 subjects have been recruited, of which disease progression has been observed in 20, and responses have been observed in only 2. One addition subject has gone through all five treatment periods without either response or progression. This subject is counted as a treatment failure. Thus, there are a total of two responses and 21 failures.

If response is treated as simple binomial variable, the posterior distribution for the response rate would be Beta(2.2, 22.8), and the predictive probability of achieving a further 50 responses out of the remaining 177 subjects would be 0.007. With such a miniscule probability of success, discontinuation of the study would be the reasonable choice.

However, examination of the distribution of outcomes over time (Table 1) calls into question the reasonableness of treating the observed outcomes as a simple binomial variable.

Table 1. Results of hypothetical study at the first futility analysis

Treatment Period	*Subjects Observed During the Period*	*Observed Disease Progressions*	*Observed Treatment Responses*
1	33	14	0
2	13	6	1
3	4	0	1
4	1	0	0
5	1	0	0
Totals	*Total subjects:* 33	*Total Progressions:* 20	*Total Responders:* 2

Of the 20 disease progressions, 14 occurred during the first treatment period, and none occurred after the second period. On the other hand, neither of the treatment responses occurred during the first treatment period. There is very little information available on later three treatment periods, and it cannot be ruled out that more information would reveal a higher responses rate during these periods. If response is treated as a simple binomial variable there is no way around this problem. However, using the method described above, the uncertainty about the later periods can be accounted for in probability calculations.

Figure 1 shows the posterior distribution of the overall number of responses if the study is allowed to run to its full planned sample size of 200 subjects. With the uncertainty with regard to the later treatment periods correctly accounted for, the predicted probability of success is much higher – 0.12 rather than 0.007. Thus, a better-informed decision can be made with regard to discontinuing the study.

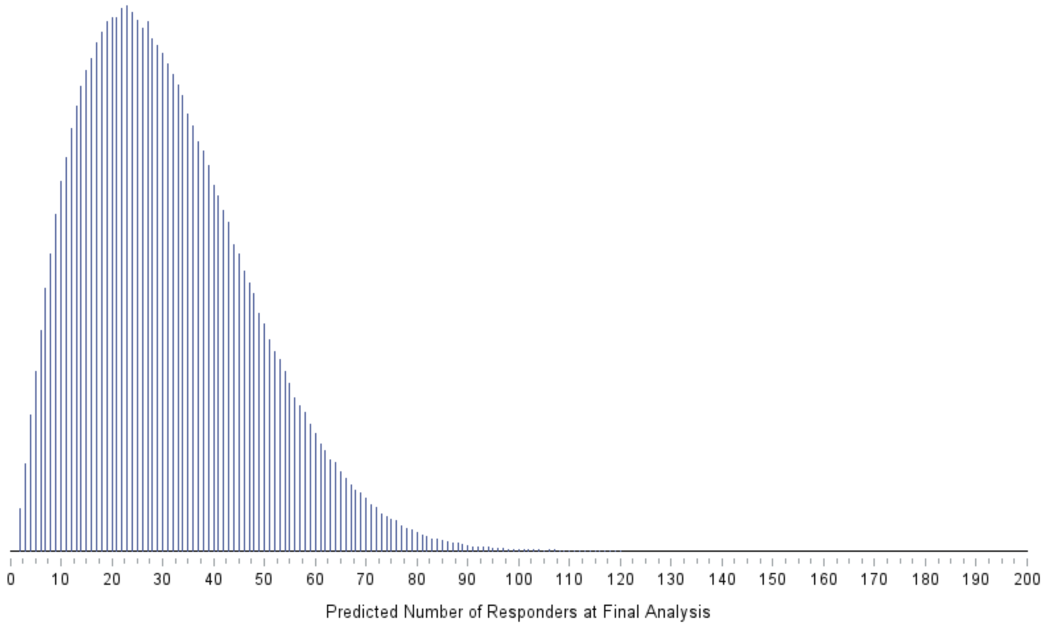

Figure 1: Probability distribution of the total number of responses if the hypothetical study is allowed to its full planned sample size of 200 subjects, given the 33 initially observed subjects summarized in Table 1, based on a Beta(0.2, 1.8) prior and response and progression probabilities each partitioned equally among the treatment periods. Based on 10,000 draws from, first, the Dirichlet distribution based on observed responses and, second, the resulting multinomial distribution.

Acknowledgements

The author wishes to thank Don Berry, whose comments helped in the development of these ideas and Chunzhang Wu and other members of the Astellas Immunology, Transplant, Central Nervous System, and Infectious Disease Statistics Group, who made several helpful comments during the development of my JSM talk.

References

Thall, P. F., and R. Simon. 1994. Practical Bayesian guidelines for phase IIB clinical trials. Biometrics 50: 337-349.