Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Not Too Big, Not Too Small: A Goldilocks Approach To Sample Size Selection

Kristine R. Broglio[a], Jason T. Connor[ab] & Scott M. Berry[a]
[a] Berry Consultants, LLC, Austin, Texas, USA
[b] University of Central Florida, Orlando, Florida, USA
Accepted author version posted online: 03 Apr 2014.Published online: 15 Apr 2014.

To cite this article: Kristine R. Broglio, Jason T. Connor & Scott M. Berry (2014) Not Too Big, Not Too Small: A Goldilocks Approach To Sample Size Selection, Journal of Biopharmaceutical Statistics, 24:3, 685-705, DOI: 10.1080/10543406.2014.888569

To link to this article: http://dx.doi.org/10.1080/10543406.2014.888569

Journal of Biopharmaceutical Statistics, 24: 685–705, 2014

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2014.888569

Taylor & Francis
Taylor & Francis Group

NOT TOO BIG, NOT TOO SMALL: A GOLDILOCKS APPROACH TO SAMPLE SIZE SELECTION

Kristine R. Broglio[1], Jason T. Connor[1,2], and Scott M. Berry[1]

[1]*Berry Consultants, LLC, Austin, Texas, USA*
[2]*University of Central Florida, Orlando, Florida, USA*

We present a Bayesian adaptive design for a confirmatory trial to select a trial's sample size based on accumulating data. During accrual, frequent sample size selection analyses are made and predictive probabilities are used to determine whether the current sample size is sufficient or whether continuing accrual would be futile. The algorithm explicitly accounts for complete follow-up of all patients before the primary analysis is conducted. We refer to this as a Goldilocks trial design, as it is constantly asking the question, "Is the sample size too big, too small, or just right?" We describe the adaptive sample size algorithm, describe how the design parameters should be chosen, and show examples for dichotomous and time-to-event endpoints.

Key Words: Bayesian adaptive trial design; Predictive probabilities; Sample size; Sequential design.

1. INTRODUCTION

A typical sample size calculation for a randomized clinical trial involves a statement of the number of patients required to achieve a desired level of power. This calculation is made conditional on an assumed control or placebo rate, a treatment effect, and a particular type I error rate. The assumed control rate may have been carefully derived from the literature, other similar trials, or even preliminary data. The treatment effect is similarly estimated, although frequently there are no published data and less preliminary data.

However, after the trial has been conducted and the data are analyzed, planned power is relatively unimportant. The actual observed treatment effect is principle. At this point either a trial has or has not achieved its objective. If the results are highly statistically significant, then the trial may have used too many resources, taken too long, and delayed conveying important information to clinicians and patients. If the results are not clinically convincing or close to statistically significant, then the last patients enrolled in the trial offered little towards providing the final evidence to achieve the trial's goal and also could have been saved (Berry, 2004).

Thus, it is very desirable during the conduct of a trial to have planned interim analyses of the accumulating data such that, if the treatment effect is large and a superiority claim is highly likely, or conversely, if the treatment effect is non-existent and futility is nearly certain, the trial can be stopped early. Group sequential techniques are widely used for this

Received July 6, 2012; Accepted March 22, 2013
Address correspondence to Kristine R. Broglio, Berry Consultants, LLC, 4301 Westbank Drive, Suite 140, Bldg B, Austin, TX 78746, USA; E-mail: kristine@berryconsultants.com

purpose (FDA, 2010). However, these methods have been criticized for being too conservative and stopping only when results are extreme (Berry, 2004). In addition, if accrual is rapid relative to when the primary endpoint is observed, accrual may be complete before enough information has accumulated for the first planned interim analysis. In this case, trial time may be saved, but not patient resources.

We present an alternative sequential design, a Bayesian adaptive sample size algorithm for a confirmatory (Phase III) trial to select a trial's sample size based on accumulating data. This design allows not just for early stopping of the trial, but for early stopping of *accrual,* and explicitly accounts for complete follow-up of all enrolled patients. Minimum and maximum sample sizes are prespecified, and frequent sample size selection analyses are made using predictive probabilities to determine whether the current sample size, if allowed to complete follow-up, is sufficient for superiority or whether continuing accrual would be futile. We refer to this as a Goldilocks design. We describe the Goldilocks adaptive sample size algorithm and show example designs for both binary and time to event endpoints, though the method may be generalized to any analysis of any endpoint or even multiple endpoints.

2. GOLDILOCKS ADAPTIVE SAMPLE SIZE ALGORITHM

Predictive probabilities have a simple interpretation that is easily communicated to clinicians, chiefly the chance of trial success if the trial continues (Dmitrienko, 2006) and as such are both natural and useful in interim assessments of ongoing clinical trials (Berry, 2004). Early stopping of clinical trials based on predictive probability calculations, particularly in the case of futility, has been described elsewhere (Berry, 1989; Choi, 1989; Dmitrienko, 2006; Herson, 1979; Julian, 2008; Kelly, 2012; Saville, in press; Spiegelhalter, 1986). In brief, for early futility stopping, the idea is to determine the probability of trial success at the final analysis if the trial were to continue to the prespecified maximum sample size. If this probability is low, the trial is stopped early for futility. The final analysis may be either a standard frequentist test or may be a fully Bayesian analysis. The predictive probability calculation includes two sources of variability: the natural variability in the data that has not yet been observed and the variability around the estimate of the treatment effect. Predictive probabilities average the trial's probability of success over the posterior distribution of the treatment effect, thus considering the knowledge of the treatment effect that has amassed and also formally accounting for the uncertainty around the treatment effect (Berry, 2004; Choi, 1989; Spiegelhalter, 1986).

The Goldilocks design allows not just for early futility stopping but also for early stopping for predicted success. The adaptive algorithm applies only during the accrual phase of the trial and so is distinct in that it is not necessarily a method for interim monitoring, but is a method for selecting a trial's sample size. Our Goldilocks adaptive sample size algorithm is based on two predictive probabilities, the predictive probability of trial success at the current sample size, P_n, if the trial were to stop accrual immediately and all enrolled patients were allowed to complete follow-up, and the predictive probability of trial success should the trial continue to the maximum sample size, P_{max}. The minimum and maximum sample size, the frequency of sample size selection analyses, the criteria for stopping accrual early for success stopping (S_n), and criteria for stopping the trial early for futility (F_n) are prespecified parameters. The trial's actual sample size is selected by comparing the two predictive probabilities to the early stopping thresholds. P_n is compared

to S_n and P_{max} is compared to F_n. S_n and F_n may be constants or may be allowed to vary across the course of the trial.

At each sample size selection analysis with n patients, if the predictive probability of trial success with the currently enrolled patients, P_n, is greater than the criteria for early success stopping, S_n,

$$P_n > S_n,$$

then the current sample size is considered highly probable to achieve success at the primary efficacy analysis. Accrual is stopped early at the current sample size, all patients are allowed to complete follow-up, and the primary efficacy analysis is conducted on those n patients. For instance if S_n is chosen to be 0.90, then stopping accrual at n patients is the equivalent of using a trial's early data to power the study at 90% rather than relying on pilot data, expert opinion, and optimism.

If the predictive probability of success at the maximum sample size, P_{max}, is lower than the criteria for early futility stopping, F_n,

$$P_{max} < F_n,$$

the trial is stopped early for futility. The logic here is that we should stop the trial immediately if expending the rest of the available resources (enrolling additional patients and additional follow-up for enrolled patients) is unlikely to produce a trial that achieves its primary aim. Otherwise, if neither early success nor futility criteria are satisfied, accrual continues and additional information is gathered for the next sample size selection analysis.

Accrual should not be paused for the sample size selection analyses; therefore, if a bound is crossed, there may be slightly more than n patients in the trial when accrual is stopped. Furthermore, the Data Safety Monitoring Board (DSMB) need not meet for each sample size selection analysis. Rather, this information should be conveyed to either the DSMB or another unblinded party with the authority to stop accrual.

While there are many sample size selection analyses, we emphasize there is a single final analysis of the primary endpoint. And thus, we also emphasize the difference between stopping *accrual* early in the case of predicted success and stopping the *trial* early in the case of futility. If accrual stops early for predicted success or if the trial continues to the maximum sample size, the primary analysis is conducted when all enrolled patients have completed follow-up for the primary endpoint. In the dichotomous endpoint case, this is when all patients have been assessed for the primary endpoint. In the time to event (TTE) case, there is a specified period of follow-up after accrual is complete, for example, 6 months or 1 year, thus ensuring that each patient either produces a minimum amount of exposure or experiences an event, and the primary analysis is conducted at the completion of this follow-up period. Thus, we may stop *accrual* early based on the strength of early evidence, which is allowed to develop with complete follow-up of the currently enrolled patients. In the case of futility, the *trial* is stopped, meaning that both accrual is halted and patients are no longer followed. This information would be conveyed to investigators, and patients would be expected to cross to a therapy that has an expectation of benefit.

At each sample size selection analysis, there will be patients with complete information (i.e., those for whom the primary endpoint is known) and more recently enrolled patients with incomplete information (i.e., those for whom the primary endpoint is not yet

available). We use patients with complete information (and partial information in the time-to-event case) to predict the outcomes for patients with incomplete information. We use Bayesian models that learn from the accruing information on the primary endpoint. Thus, there is more uncertainty around the predictions early in the trial than later in the trial when more primary endpoint data has been amassed. For example, when calculating P_{max} at early analyses, there may be fewer patients offering complete data than there are outstanding patients, whereas at later analyses the majority of enrolled patients may offer complete data and therefore more observed data are being used to predict outcomes for the few outstanding patients.

In this setting, accrual rate and the time required to observe the primary endpoint are important design elements. This adaptive design will function best when accrual of patients is slower relative to observing the primary endpoint. At an extreme, if accrual is very fast and time to the primary endpoint is very long, there could be no primary endpoint information available at each sample size selection analysis, and thus a large amount of uncertainty will remain in the predictive probability calculations, hindering the ability of the algorithm to make the sample size selection decisions. As time to the primary endpoint grows shorter and the ratio of available data to data to be predicted grows larger, P_n will converge to either 0 or 1. Because there is little to no outstanding data, the predictive probability of trial success with the current sample size either does or does not achieve its prespecified goal. Thus, if the primary endpoint is known nearly instantly, such as within hours or days, early success monitoring could be based on posterior probabilities while futility monitoring remains based on predictive probabilities.

3. DICHOTOMOUS ENDPOINT

As an example, imagine a two-arm blinded randomized controlled clinical trial with a dichotomous endpoint observed at 60 days. A traditional chi-squared test will be used to compare groups.

3.1. Beta-Binomial Model

In this section we refer to the dichotomous endpoint generically as success or failure. We assume the probability of success, θ_j, has a Beta prior distribution:

$$\left[\theta_j\right] \sim \text{Beta}\left(\alpha, \beta\right),$$

where $j = c$ is the control arm and $j = e$ is the experimental arm. At each sample size selection analysis the number of observed successes, x_j, among the currently enrolled patients, n_j, follows a binomial distribution:

$$\left[x_j\right] \sim \text{Binomial}\left(n_j, \theta_j\right).$$

We update the prior distribution with the currently observed data (x_j, n_j) and the resulting posterior distribution (Gelman, 2003) is

$$\left[\theta_j | x_j, n_j\right] \sim \text{Beta}\left(\alpha + x_j, \beta + n_j - x_j\right).$$

Here, we assign the same vague prior to both the control and experimental group, Beta(α,β). Informative priors could be selected and/or the priors could be allowed to be different for the two arms. The prior is used, along with accumulating data from the trial, only to predict outstanding data. The chi-squared test is used to compare groups at the end of the trial. Therefore, using different priors for the control and experimental groups will only affect when to terminate accrual and not the primary analysis. If the prior is skeptical (e.g., the treatment will perform more poorly than the control) the predictive probability of success will be lower and the trial will be less likely to stop early for predicted success. More overwhelmingly positive data will be necessary to stop the trial. Likewise, if an optimistic prior is used (e.g., the treatment will perform better than the control) then the trial may stop accrual for predicted success too easily and the final result will not reach statistical significance. This is equivalent to a fixed trial with an overly optimistic treatment effect chosen during the design stage—the actual power is lower than anticipated.

Determining the prior is analogous to estimating the control rates and treatment effects in the design stage of any standard trial. Only in this design, we apply more variability than point estimates of the control rates and treatment effects and allow our prior beliefs to be updated by accumulating information from within the trial.

3.2. Predictive Probability of Success at Current Sample Size

Predictive probabilities of trial success are calculated using two beta-binomial distributions, one for the control group and one for the experimental group. Groups are modeled independently.

At each sample size selection analysis with a total of n_j patients enrolled, the number of patients with a known outcome and who have achieved success is x_j and the number with a known outcome who have not achieved success is z_j. The number of currently enrolled patients with an unknown outcome is n^*_j. Thus, $n_j = x_j + z_j + n^*_j$. Among the n^*_j patients, we denote x^*_j as the number who will ultimately achieve success. x^*_j is a random variable distributed beta-binomial,

$$x^*_j | x_j, z_j, n^*_j \sim \text{Beta} - \text{binomial}(n^*_j, \alpha + x_j, \beta + z_j).$$

Each possible value of x^*_j has an associated probability, $\Pr(x^*_j)$, based upon the beta-binomial distribution described already and given the observed x_j and z_j. Likewise, every pair of data $(x_C + x^*_C,\ x_E + x^*_E)$ can be applied to the chi-squared test to see whether it achieves or fails to achieve the prespecified critical value for the hypothesis test.

The predictive probability of success at the current sample size, n, is then

$$P_n = \sum_{x^*_E=0}^{n^*_E} \sum_{x^*_C=0}^{n^*_C} I\left(x^*_E, x^*_C\right) \Pr\left(x^*_E\right) \Pr\left(x^*_C\right),$$

where $I(x^*_E,\ x^*_C)$ is an indicator function that equals 1 for values of x^*_E and x^*_C that would result in trial success (the chi-squared test producing a p-value below the critical value) and equals 0 for values of x^*_E and x^*_C that would result in trial failure.

3.3. Predictive Probability of Success at the Maximum Sample Size

For the predictive probability of success at the maximum sample size, we perform a similar calculation as described already, but now assume the trial will continue to the maximum sample size. This requires calculation of the predictive distribution including patients who are not yet enrolled. Hence, in

$$x^*_j|x_j, z_j, n^*_j \sim \text{Beta} - \text{binomial}(n^*_j, \alpha + x_j, \beta + z_j),$$

n^*_j is updated to include both patients who are enrolled but have not yet provided complete outcomes and the $N_{max}/2 - n_j$ patients who are yet to be enrolled to group j.

This calculation then proceeds as described earlier. We denote P_{max} as the predictive probability of trial success if the trial continues to the maximum sample size.

4. TIME-TO-EVENT ENDPOINT

We can also apply this design to a trial with a time-to-event endpoint. The final analysis could be a log-rank test, a Cox proportional hazards model, or a fully Bayesian model where the posterior distributions of the hazard rates are compared between groups. Let us assume a Cox proportional hazard model is prespecified with the inference on the hazard ratio for treatment.

4.1. Gamma Exponential Model

At each sample size selection analysis currently enrolled subjects who have experienced the event of interest contribute complete information to the Cox model that will be used in the final analysis. Enrolled patients who have not yet experienced the event of interest contribute exposure time to the Cox model; however, whether and when these patients will experience an event remains uncertain. Therefore, we calculate a predictive distribution for each patient's future event time using simulation.

We assume exponential survival times and that the hazard rate on each arm λ_j has a Gamma prior distribution:

$$\lambda_j \sim \Gamma(\alpha, \beta),$$

where α represents the prior number of events and β the prior exposure time in group j. Therefore, α/β is the prior mean for the rate, λ_j.

At each sample size selection analysis with n_j enrolled patients in group j, the total number of observed events is EV_j and the total observed exposure time is EXP_j per group, assuming each patient i in group j offers $\text{EXP}_{i,j}$ exposure. We update the prior distribution with the currently observed data (EV_j, EXP_j) and the resulting posterior distribution (Gelman, 2003) is

$$\lambda_j \sim \Gamma(\alpha + EV_j, \beta + EXP_j).$$

4.2. Predictive Probability of Success at Current Sample Size

The predictive probabilities of trial success in the TTE case are calculated using Monte Carlo integration. For the predictive probability of trial success at the current sample size, we assume accrual is stopped and the primary analysis is conducted after completion of a prespecified amount of follow-up.

A single hazard rate, λ_j, is sampled from the posterior distribution. For each patient who has not yet experienced an event, we sample an event time from an exponential distribution with mean equal to the sampled hazard rate,

$$TTE_{i,j} \sim \text{Exponential}(\lambda_j).$$

Total trial duration is the accrual period plus a follow-up period. Assuming accrual is stopped at the current sample size, all enrolled patients have the same remaining follow-up period, for example, 6 months. If the sampled event time occurs before the end of this follow-up, the event is considered observed at $TTE_{i,j}$. For each patient i with a simulated observed event, his or her exposure time is $EXP_{i,j} + TTE_{i,j}$. Otherwise, the patient is considered censored at the end of the trial, so his or her additional exposure time is $EXP_{i,j} +$ *follow-up* (e.g., 6 months).

The final Cox proportional hazard analysis is conducted on the simulated complete data set, and success or failure, as determined, for example, by comparing the p-value for the treatment's hazard ratio to 0.05, is recorded. This process is repeated 100,000 times, and the proportion of simulations in which trial success is achieved is the predictive probability of trial success at the current sample size, P_n.

4.3. Predictive Probability of Success at the Maximum Sample Size

For the predictive probability of success at the maximum sample size, P_{max}, we additionally consider accrual of each of the future patients to be enrolled to achieve the maximum sample size. This scenario has the added complexity that if follow-up is a fixed amount of time after the final patient is enrolled, then each patient yet to be accrued has a different follow-up time. We assume an accrual rate for the future patients. This assumption would be prespecified and could be, for example, the average of the accrual rate for the trial, or the average accrual rate in the last 6 months. The total trial duration is the total accrual time plus the follow-up period. Accrual times of future patients are simulated from a Poisson process using the assumed accrual rate. Event times for future patients are simulated as already described. If the event occurs before the end of the study, the event is considered observed. Otherwise, the patient is considered censored at the end of the trial. The final Cox proportional hazard analysis is conducted with the complete data set. This is repeated 100,000 times and the proportion of simulations that result in trial success represents the predictive probability of trial success at the maximum sample size, P_{max}.

5. SAMPLE SIZE CONSIDERATIONS

Our Goldilocks trial design requires prespecification of several parameters, including the minimum and maximum sample size; early stopping probabilities; number, frequency, and timing of the sample size selection analyses; and in the TTE case the minimum follow-up period after the final patient is enrolled. All these design parameters work in concert

with each other and must be tuned to optimize the design's performance. In addition, as with traditional group sequential methods, the alpha level of the final analysis may need to be adjusted to accommodate the sample size selection analyses and preserve the trial's overall type I error rate. Thus, each Goldilocks trial design requires extensive trial simulation under multiple null hypothesis scenarios, where there is no difference between treatment and control, and under various alternative hypothesis scenarios, where there are varying treatment effects.

The minimum sample size can be chosen clinically, as the minimum number of patients the trial team or key stakeholders would be comfortable with before making a decision to stop the trial or the minimum number of patients needed for a safety assessment. Through simulation, if the trial appears highly unlikely to stop at the first sample size selection analysis because of little accumulated primary endpoint information and large uncertainty around the treatment effect, the minimum sample size may be increased. Reducing the number of sample size selection analyses may allow for a higher alpha level at the final analysis.

The maximum sample size can be chosen financially or operationally, as the maximum number of patients that the sponsor can afford, that can be reasonably enrolled in a certain period of time, or, more ideally, chosen statistically as the sample size that provides an adequate probability of success for the expected treatment effect. The first choice for the maximum sample size may be that from a standard sample size calculation assuming a fixed trial design. The suitability of the maximum sample size can be assessed by simulation. If simulation results of the alternative hypothesis scenarios show a high probability of success and a high probability of stopping accrual early, the maximum sample size can be decreased. Alternatively, if simulations show the trial most certainly continuing to the maximum sample size and having an insufficient probability of success, the maximum sample size should be increased.

Early stopping thresholds, S_n and F_n, are typically selected within certain ranges. For example S_n is typically 80–99% and F_n is typically 2.5–20%. S_n and F_n may be held constant throughout the trial, or may vary by sample size selection analysis such that a higher threshold of certainty is required for stopping at smaller sample sizes and this is lowered later in the trial when more data has accrued. These values are tuned by considering several different simulation results. First, if the simulated trials are rarely stopping for futility in the null hypothesis scenario then F_n should be increased, or if simulated trials are rarely stopping accrual early for predicted success in optimistic alternative hypothesis scenarios then S_n should be decreased, making it easier to stop early. Second, in the case of stopping accrual early for success, if simulated trials that stop accrual early for success are failing to be successful at the final analysis, the S_n criterion should be increased. In the case of stopping early for futility, if simulated trials that stop early would have been successful had they been allowed to continue to the maximum sample size, the F_n criterion should be decreased. Some futility stopping under the alternative hypothesis is expected. For example, in a standard trial with standard type II errors of 10–20% it is better to have the negative trial stop sooner (via futility stopping) than later (at the maximum sample size).

Finally, the frequency of sample size selection analyses must be defined by either the number of patients enrolled, after every 50 patients are enrolled, for example, or by calendar time, every 4 weeks, for example. The frequency of the sample size selection analyses is tied to the accrual rate and is, in part, an operational decision, but should be conducted as often as feasible. Sample size selection analyses require only the primary endpoint data elements, so conducting these analyses is a much more streamlined process as compared to

a comprehensive interim safety analyses that may be conducted for a DSMB. Many sample size selection analyses will increase the trial's ability to respond to accruing data, but will also slightly increase the probability of a type I error.

Trial results can be categorized as (1) stopping for predicted success and achieving success at the final analysis, (2) stopping for predicted success and not achieving success at the final analysis, (3) stopping early for futility, (4) continuing to the maximum sample size and achieving success, and (5) continuing to the maximum sample size and not achieving success. Under the null hypothesis, type I error will be the sum of the probability of stopping early for predicted success and achieving success and continuing to the maximum sample size and achieving success. Under the alternative hypothesis, this sum is the trial's power. For all but the simplest, typically dichotomous, cases, simulation is necessary to calculate the operating characteristics, including the overall type I error rate, of the trial.

The null hypothesis is that the control rate is equal to the treatment rate and the null space is completely defined by this treatment effect, the control rate, and the accrual rate. Type I error should be controlled across the entire null space and therefore across the ranges of all reasonably likely control rates and accrual rates. Because type I error is simulation based, it is also important to perform tens of thousands of simulations of each null scenario to reduce simulation error in these calculations.

The type I error rate should be extensively explored in the simulation process and trade-offs between parameters evaluated in order to control type I error at the desired level, such as less than 5%. As the number of sample size selection analyses increases, the alpha level at the final analysis may also need to be more stringent. With a high hurdle for success at the final analysis, it will be more difficult to stop accrual early for success and more likely to predict futility, so S_n and F_n may need to be relaxed to achieve desirable results.

6. EXAMPLE TRIAL: GENESEARCH BREAST LYMPH NODE ASSAY

Julian et al. (2008) provide an example of a Goldilocks trial design with a dichotomous endpoint. This was a single-arm study of the GeneSearch Breast Lymph Node (BLN) Assay for patients with breast cancer undergoing sentinel lymph node (SLN) biopsy. The primary objective of the study was to characterize the test's sensitivity and specificity. The gold standard was histologic evaluation. A positive/negative result for metastasis in the SLN was available intraoperatively. Because of the immediate nature of the primary endpoint, this trial used posterior probabilities for monitoring of success at the current sample size and predictive probabilities for monitoring of futility at the maximum sample size.

Sample size selection analyses were planned starting when 200 patients were enrolled and after every additional 50 patients. The predictive probability of trial success is based upon achieving success for two endpoints. Because data availability was essentially immediate, the trial would stop early for success if, based on the posterior distributions of the sensitivity (π_1) and specificity (π_0),

$$\Pr(\pi_1 > 0.70 \mid \text{Data}) \geq 0.985 \text{ and } \Pr(\pi_0 > 0.90 \mid \text{Data}) \geq 0.985.$$

The trial would stop early for futility if the predictive probability of trial success at the maximum sample size of 700 patients was less than 5% (F_n). Simulation of the trial design showed that these criteria controlled the trial's overall type I error rate to less than 5%. R code (R Core Development Team, 2011) to simulate this trial design is provided in the appendix.

Accrual was stopped early for success after 416 patients had been enrolled. The final observed sensitivity was 87.6% (95% confidence Interval = 80.4%, 92.9%), and the final observed specificity was 94.2% (95% confidence Interval = 90.9%, 96.6).

7. EXAMPLE TRIAL: TIME TO EVENT IN ONCOLOGY

This is a generic example of a Goldilocks trial design with a time-to-event endpoint. The setting is a two-arm randomized trial in oncology where patients are equally allocated to either a control or a treatment arm. The primary endpoint is overall survival (OS) measured from the date of randomization to the date of death from any cause or last follow-up. The expected OS rate at 1 year for the control arm is 30%. The minimum sample size is 100 and the maximum sample size is 300. The follow-up period after accrual is complete is 1 year. Thus, if accrual is stopped early for predicted success or the trial continues accrual to the maximum sample size of 300 patients, the primary analysis of OS will be conducted after an additional 1 year of follow-up for all patients.

Design parameters were selected through the process of simulating and examining the trial's operating characteristics as described above. Sample size selection analyses are planned starting when 100 patients are enrolled and after every additional 25 patients are enrolled. Early stopping for futility is allowed starting with the 100 patient sample size selection analysis and F_n is 10%. Stopping accrual early for predicted success is only allowed starting with the 200 patient sample size selection analysis and S_n is 90%. Based on the simulation of the type I error, in order to control the overall one-sided type I error at 2.5% considering 1-year OS rates for the control arm from 10% to 50% and accrual rates in the range of 5 to 10 patients per month, the final analysis of OS is based on a one-sided log-rank test at the $\alpha = 0.022$ level.

Table 1 shows selected operating characteristics of this trial including the average sample size, probabilities of early stopping, and probability of trial success under varying assumptions of the hazard ratio (HR). Results assume OS at one year is 30% for the control arm and that patients are accrued at the rate of 5 per month. Type I error is demonstrated by the probability of a successful trial when there is no difference between the arms, when the HR is 1.0. In this scenario, the probability of trial success is 0.023, and the trial stops early for futility with 0.93 probability. The mean sample size is 153, indicating the trial is likely to stop between the 100 and 200 patient sample size selection analyses. Trial power is

Table 1 Example operating characteristics assuming accrual of 5 patients per month

Control 1-year survival	Treatment 1-year survival	Hazard ratio	Mean (SD) sample size	Stop for futility	Stop for max N	Stop for pred succ	Probability of success
0.30	0.30	1.00	152.8 (62.1)	0.93	0.06	0.02 (0.002)	0.023
0.30	0.38	0.80	215.4 (70.4)	0.46	0.26	0.29 (0.018)	0.39
0.30	0.43	0.70	223.8 (54.5)	0.18	0.20	0.62 (0.016)	0.75
0.30	0.49	0.60	209.5 (32.3)	0.04	0.05	0.91 (0.005)	0.95

Note. Parenthetical values below proportion of trials that stop for predicted success (pred succ) are the values that then fail to reach statistical significance at the final analysis.

Table 2 Example operating characteristics assuming accrual of 10 patients per month

Control 1-year survival	Treatment 1-year survival	Hazard ratio	Mean (SD) sample size	Stop for futility	Stop for max *N*	Stop for pred succ	Probability of success
0.30	0.30	1.00	167.2 (68.8)	0.88	0.10	0.02 (0.006)	0.024
0.30	0.38	0.80	223.5 (73,1)	0.44	0.34	0.22 (0.025)	0.35
0.30	0.43	0.70	230.8 (60.4)	0.20	0.29	0.51 (0.023)	0.70
0.30	0.49	0.60	219.7 (41.9)	0.06	0.11	0.83 (0.007)	0.93

Note. Parenthetical values below proportion of trials that stop for predicted success (pred succ) are the values that then fail to reach statistical significance at the final analysis.

demonstrated by probability of success in scenarios where the HR is less than 1. As the HR decreases, the trial is able to stop for predicted success more frequently and the probability of a successful trial increases.

Accrual rate is an important element in these trial designs. When events are observed only after some follow-up time has elapsed, slower accrual allows for more information available at each sample size selection analysis. Table 2 shows the operating characteristics assuming an accrual rate of 10 patients per month. The fast accrual results in a decreased ability to stop accrual early and thus higher mean sample sizes. As accrual becomes more rapid, the total exposure time and number of observed events will decrease, so the trial also has a decreased probability of success. We also report the proportion of trials that stop for predicted success and fail to reach statistical significance. The probability of stopping for predicted success but failing the primary analysis when the HR is 0.6 is 0.5% when accrual is 5 patients per month and 0.7% when accrual is 10 patients per month.

Figure 1 shows the probability of trial success (left panel) and the mean sample size (right panel) for this example Goldilocks design and a traditional group sequential design. For the traditional group sequential design we used the gamma family spending function with a parameter value of –4 for both the upper and lower bounds. For comparison purposes, we assumed a maximum sample size of 300 patients, 95% power, a one-sided type I error rate of 2.5%, and the same number of interim analyses, though we equally spaced them across the expected information. This group sequential design required a maximum of 203 events. The required one-sided p-value at the final analysis is 0.0199.

The Goldilocks design tends to have a higher probability of success as compared to the traditional group sequential design (Fig. 1, left panel). By design, both have a 95% probability of success when the true HR is 0.60 and no greater than a 2.5% probability of success when the true HR is 1. For more moderate hazard ratios, the Goldilocks design has approximately 5–10% higher probability of success.

At the accrual rate of 5 patients per month, the Goldilocks design does tend to have larger mean sample sizes for the more moderate hazard ratios, the trade-off for a high probability of success. At the accrual rate of 10 patients per month, the Goldilocks design achieves a higher probability of success for the more moderate hazard ratios than does the traditional group sequential design at a lower sample size. The Goldilocks design is based on of the number of patients accrued rather than the number of events observed. As accrual becomes faster relative to observed events, the Goldilocks design is able to stop accrual for predicted success and continue follow-up of patients in order to observe the

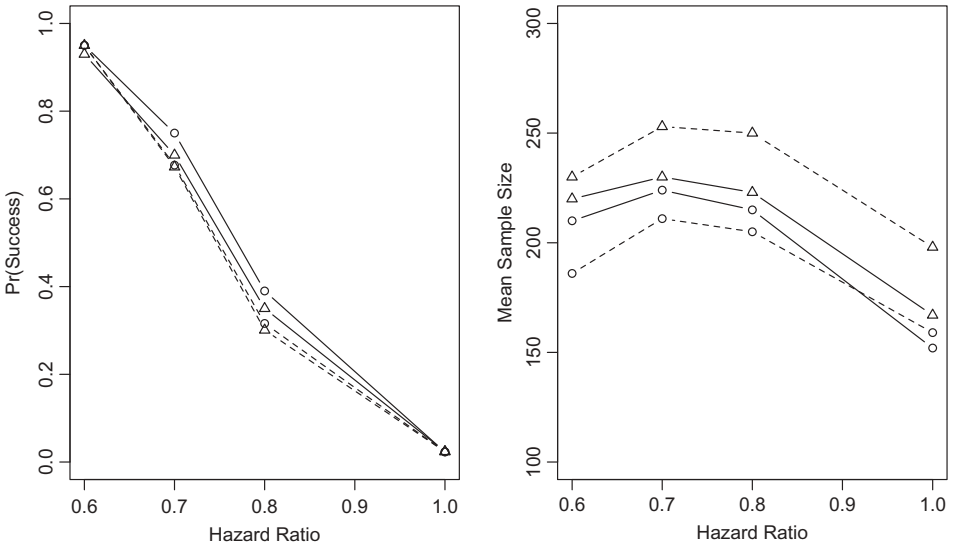

Figure 1 Probability of trial success (left) and mean sample size (right) for the Goldilocks TTE example and a traditional group sequential design. Solid lines are the Goldilocks design and dashed lines are the traditional design. Circles represent an accrual rate of 5 patients per month, and triangles represent an accrual rate of 10 patients per month.

necessary number of events, whereas the traditional group sequential design is only able to stop accrual once the necessary events have been observed. In some settings, with a traditional group sequential design, accrual could be complete before enough events have been observed to stop a trial early.

We also use this example to illustrate how the accrual rate and the follow-up time affects type I error. Table 3 shows type I error for 4 accrual rates (2, 5, 8, and 15 patients per month) and 4 follow-up times (1, 6, 12, and 24 months) assuming 1 year OS in both arms is 30%. These are based on 10,000 simulations per scenario. The standard error due to simulation is 0.0016 (∼0.16%). Slower accrual rates and shorter follow-up times after the end of accrual spend more type I error rate. The less outstanding data (patients with unobserved events), the more certain are the predictions; therefore, random highs leading to accrual stopping will be less likely to regress to the mean. However, with longer follow-up times there is more time for regression to the mean and random highs are no

Table 3 Type 1 error rates by accrual rate and follow-up time, assuming 1 year OS of 30% for control and treatment

Accrual rate, patients/month	Follow-up time after last patient enrolled (months)			
	1	6	12	24
2	0.032	0.028	0.025	0.023
5	0.029	0.025	0.023	0.022
8	0.028	0.022	0.022	0.019
15	0.026	0.022	0.020	0.018

longer statistically significant at the end of follow-up. If accrual is expected to be slow, then a greater type I error adjustment would be necessary. Here even with slow accrual (unfeasibly slow in that the trial could take 12.5 years with 2-patient-per-month accrual) the type I error is controlled for 12-month follow-up, the actual follow-up in this trial.

8. CONCLUSIONS

We present a Bayesian adaptive sample size algorithm, a Goldilocks trial design, for optimizing the sample size of a Phase III or confirmatory trial within a prespecified range. We show models and examples for dichotomous and time-to-event endpoints, though the algorithm applies more generally and can be used regardless of the primary efficacy analysis. We focused on superiority trials, though the algorithm extends to noninferiority trials as well.

The Goldilocks design is a sequential design that offers many advantages. It is ideal for responding to accumulating data during accrual, especially in settings where there is a small to moderate delay between enrolling patients and observing their primary outcome, and this design has been successfully implemented in several settings in all three branches of the Food and Drug Administration (FDA), CDER (White, 2012), CBER (Oncolytics Biotech, n.d.), and CDRH (AtriCure, n.d.; Jullian, 2008; Wilbur, 2010), as well as by the National Cancer Institute (NCI)-sponsored CALGB (Muss 2009). It has been the design used leading to multiple device approvals, including but not limited to AtriCure (n.d.), Julian (2008), and Wilbur (2010). While the Goldilocks design shares some of the same ideals as a group sequential design, there are some important differences. Interim analyses in the Goldilocks design are based on the number of patients enrolled rather than the amount of observed information, and complete follow-up of all enrolled patients is an explicit part of the design. While a group sequential design would stop immediately upon crossing a success bound, the Goldilocks design would stop accrual but allow for complete follow-up of all patients. Thus, with a group sequential design, the last patient enrolled might contribute very little to the inference. Furthermore, with a traditional group sequential design, it may not be straightforward or prespecified how to analyze and report the trial results once all enrolled patients have completed their assigned treatment and follow-up. Hampson and Jennison (2013) illustrate a case of using group sequential design with delayed responses. In the Goldilocks design, each patient enrolled contributes fully, by design and in a prespecified manner, to the statistical inference. Allowing for complete follow-up means that, under the alternative hypothesis, the mean sample size may be smaller. Under the null, even if the early data were promising enough to stop accrual early, we would expect regression to mean during follow-up and to not attain success at the final analysis, thus preventing a type I error. As such, less type I error is spent at each sample size selection analysis. Results need not already be highly convincing for success (or futility) in order to stop the trial early—only that the early data suggest a treatment effect such that, if the trend were allowed to play out, we are convinced that success (or futility) would be declared.

Furthermore, the Goldilocks design is incredibly flexible. Any endpoint and any final analysis, Bayesian or frequentist, can be accommodated. The early stopping values of S_n and F_n can be calibrated to be more or less conservative. A natural extension of the design described here is to incorporate earlier measures of the primary endpoint into the predictive probability calculations. This improves the predictive ability of the algorithm and extends the design's usefulness into settings where the primary endpoint may be more long term.

Unlike group sequential methods that consider endpoints individually, the predictive probability calculations can be extended to accommodate multiple endpoints, as was done in Julian (2008). The vast majority of regulatory endpoints are binary and time to event, so we do not present the continuous case. The design is generalizable to the continuous case with appropriate specification of the prior and the posterior distributions for continuous variables.

For any particular clinical trial, there is tremendous flexibility in applying the Goldilocks design, as well as a large number of well-understood group sequential alternatives. Our time-to-event example includes a comparison to a traditional group sequential trial. However, in our experience, the Goldilocks design will incorporate more frequent interim analyses than is typically for a group sequential design, particularly because less type I error may be spent at each interim analysis. When there are only two or three planned interim analyses, a traditional group sequential trial will be less likely to stop early and the difference in mean sample size will be even greater between the two approaches. If a Goldilocks design bases stopping for predicted success on posterior probabilities, then that boundary is likely to approximate one of the upper bounds in the family of group sequential designs. Still, the Goldilocks design would offer greater flexibility and the use of predictive probabilities for the futility boundary. Regardless, it is important to thoroughly evaluate any adaptive design, to understand the benefits and risks, and to compare to alternative designs (Gaydos, 2009; Patient-Centered Outcomes Research Institute [PCORI], 2012).

A common concern with any Bayesian approach is the specification of the prior distribution. All trial designs require eliciting a prior belief in the control rates and the treatment effects, and designs are based on these assumptions, whether they are written down in the form of a prior distribution or not. If the primary analysis is frequentist, the prior distributions used in the Goldilocks design will have no bearing at all on the final analysis; they only lead to stopping being more or less aggressive. If a fully Bayesian approach is specified, noninformative priors that have no effect on the inference can be used. Different priors may be specified for use in the adaptive sample size algorithm than for the primary analysis. In this manner, informative priors could be specified for the adaptive sample size algorithm, but noninformative priors could be used for the primary analysis. If informative priors are deemed appropriate, the influence of the priors should be thoroughly evaluated at the design stage. This is analogous to any standard frequentist trial: A point prior in the alternative hypothesis space is chosen during the design stage under which the power is calculated, and then a noninformative prior (a standard frequentist test) is used in the analysis stage. The difference with the Bayesian approach is that our prior is updated as data from the trial accumulate, so that the final sample size need not be too onerous or does not include patients after the point where continuing the trial is futile. Then either a frequentist test or a Bayesian test with a noninformative prior is typically used in the analysis stage.

Operating characteristics, and particularly type I error control, require simulation for calculation. A benefit is that this forces designers to better understand their trials and how they are likely to progress. For example, with simulated trials we can easily calculate, under various alternative hypotheses, what proportion of trials that stop for futility would have produced successful trials if left to run to the maximum. This is a seldom-reported value in group sequential designs. Furthermore there is precedent for type I error calculation via simulation by CDER, CBER, CDRH, and NIH. Drawbacks are that the absolute maximum of the type I error rate over the null space cannot be calculated in closed form, and simulations, depending upon the trial, can be slow to run. Some adjustment of trial inputs is necessary to control type I error.

The Goldilocks design could be described as a design in which a minimum sample size is specified and the sample size is *increased* (to a prespecified maximum) based on interim analyses, or alternatively as a design in which a maximum sample size is specified and there is potential to *decrease* the sample size (to a prespecified minimum) based on interim analyses. These are equivalent statements. Conduct of the trial does rely on unblinded analyses of the treatment effect, and in that regard there is potential for the introduction of operational bias (FDA, 2010). Thus, in conduct, it is important to follow the same principles as for any other trial, and with the appropriate procedures in place the potential for bias is minimal (He, 2012; PCORI, 2012): namely, that the rules of trial conduct be completely prespecified and carefully adhered to, that interim results be kept confidential and blinding maintained, and that a DSMB be tasked with trial oversight (FDA, 2010; He, 2012; PCORI, 2012). The DSMB charter should be explicit with regard to how the results of the sample size selection analyses will be conveyed, to whom, and how the results will be acted upon.

APPENDIX

This appendix provides R code (R Development Core Team, 2011) for simulating the trial described in Julian et al. (2008).

```
library(VGAM)
#### Set design parameters
design.parameters <- list(
sens.goal   = 0.7,                          # Goal sensitivity
spec.goal   = 0.9,                          # Goal specificity
looks        = seq(200, 700, by=50),        # Frequency of analyses
sens.alpha = 0.1,                           # Priors for sensitivity
sens.beta   = 0.1,
spec.alpha = 0.1,                           # and specificity
spec.beta   = 0.1,
ref.alpha   = 0.1,                          # Prior for incidence needed
ref.beta    = 0.1,                          # to calc pred prob win @ max
sens.critv = 0.985,                         # Critical values for trial
success
spec.critv = 0.985,                         #    .985 gives 0.025 type I error
 min.hist.pos = 35,                         # Min Samples positive on
                                                 histology to stop early
fut.bound   = 0.05                          # Futility stopping bound
)
### Make a vector of the number of 'right' responses for
###  each possible total number of positives
needed.to.win.sens <- rep(NA,700)
needed.to.win.spec <- rep(NA,700)
# What's fewest test positives we can have for N histology positives?
min.right.sens <- function(N, design.parameters){
       x <- 0
       post <- 0
       while(post < design.parameters$sens.critv){
```

```
        x <- x+1
        post <- 1-pbeta(design.parameters$sens.goal, .1+x, .1+N-x)
        }
        return(x)
}
# What's fewest test negatives we can have for N histology positives?
min.right.spec <- function(N, design.parameters){
        x <- 0
        post <- 0
        while(post < design.parameters$spec.critv){
        x <- x+1
        post <- 1-pbeta(design.parameters$spec.goal, .1+x, .1+N-x)
        }
        return(x)
}
for(n in 35:700){
     needed.to.win.sens[n] <- min.right.sens(n, design.parameters)
     needed.to.win.spec[n] <- min.right.spec(n, design.parameters)
}
### FUNCTION THAT GENERATES TRIAL DATA
make.data <- function(N, true.hist.pos, true.sens, true.spec){
     # N is sample size
     # generate whether each of N patients is a reference positive
     hist.pos <- rbinom(N, 1, true.hist.pos)
     # generate whether each of N patients tests positive
     test.pos <- rbinom(N, 1, ifelse(hist.pos==1, true.sens,
     1-true.spec))
     # Combine & return
     data <- cbind(hist.pos, test.pos)
     return(data)
}
### Function that does the analysis at each interim (or final) analysis
primary.analysis <- function(N, data, design.parameters, pr){
        pm <- design.parameters
        tab2x2 <- table(data[1:N,1], data[1:N,2])
        # Hist positives in rows
        # Test positives in columns
        # Table is:
        # Hist -,  Test - Hist -, Test +
        # Hist +, Test - Hist +, Test +
        # Sensitivity = Pr(Test + | Hist +)
        # Probability exceed sensitivity goal
        PostProbSens <- 1-pbeta(pm$sens.goal, pm$sens.alpha + tab2x2[2,2],
                                              pm$sens.beta + tab2x2[2,1])
        # Specificity = Pr(Test - | Hist -)
        PostProbSpec <- 1-pbeta(pm$spec.goal, pm$spec.alpha + tab2x2[1,1],
                                              pm$spec.beta + tab2x2[1,2])
        # Predictive probability of trial success at max
```

```
        N.left <- max(pm$looks) - N
        if(N.left > 0){
        # Prob dist for number of hist positives remaining
        x.left.hist.pos <- rbetabin.ab(10000, N.left,
                              pm$ref.alpha + tab2x2[2,1] + tab2x2[2,2],
                              pm$ref.beta  + tab2x2[1,1] + tab2x2[1,2])
        x.left.hist.neg <- N.left - x.left.hist.pos

        x.left.hist.pos[x.left.hist.pos==0] <- 1 ####*
        x.left.hist.pos.test.pos <- rbetabin.ab(10000, x.left.hist.pos,
                              pm$sens.alpha + tab2x2[2,2],
                              pm$sens.beta  + tab2x2[2,1])
        x.left.hist.pos.test.pos[x.left.hist.pos==0] <- 0 ####*
           #### * these two lines are here because if it draws 0
           #### hist positives left, this gives an error.
           #### so this allows the draw, but changes it back to 0
        x.left.hist.neg.test.neg <- rbetabin.ab(10000, x.left.hist.neg,
                              pm$spec.alpha + tab2x2[1,1],
                              pm$spec.beta  + tab2x2[1,2])

        ### Is number of test positives big enough to achieve sens goal.
        hist.pos.at.max <- tab2x2[2,1]+tab2x2[2,2]+x.left.hist.pos
        hist.pos.test.pos.at.max <- tab2x2[2,2] + x.left.hist.pos.test.pos
        win.sens.at.max <- (hist.pos.test.pos.at.max >=
          needed.to.win.sens[hist.pos.at.max])
        ### Is number of test negatives big enough to achieve spec goal.
        hist.neg.at.max <- tab2x2[1,1]+tab2x2[1,2]+x.left.hist.neg
        hist.neg.test.neg.at.max <- tab2x2[1,1] + x.left.hist.neg.test.neg
        win.spec.at.max <- (hist.neg.test.neg.at.max >=
          needed.to.win.spec[hist.neg.at.max])
        pred.prob.both.goals.at.max <- mean(win.sens.at.max*win.spec.at.max)
        }else{
        pred.prob.both.goals.at.max <- NA
        }
        # Function returns     1) total N
        #                      2) number positive on histology
        #                      3) post. prob sens meets goal
        #                      4) post. prob spec meets goal
        #                      5) pred prob sens&spec meet goal at Nmax
        if(pr){print(c(N, c(tab2x2), PostProbSens, PostProbSpec,
pred.prob.both.goals.at.max))}
        return(c(N, tab2x2[2,1]+tab2x2[2,2], PostProbSens, PostProbSpec,
pred.prob.both.goals.at.max))
}
#### CHECK TO SEE IF STOPPING BOUNDARIES MET
stop.check <- function(interim.analysis, design.parameters,pr){
        go <- 1
        if(interim.analysis[2] >= design.parameters$min.hist.pos &
```

```
          interim.analysis[3] >= design.parameters$sens.critv &
          interim.analysis[4] >= design.parameters$spec.critv
        ){
         go <- 0
         win <- 1
         stop <- 3
         n.trial <- interim.analysis[1]
        }else if(interim.analysis[1] >= max(design.parameters$looks)){
         go <- 0
         win <- 0
         stop <- 2
         n.trial <- interim.analysis[1]
        }else if(interim.analysis[2] >= design.parameters$min.hist.pos &
          interim.analysis[5] <= design.parameters$fut.bound){
         go <- 0
         win <- 0
         stop <- 1
         n.trial <- interim.analysis[1]
        }else {
         go <- 1
         win <- NA
         stop <- NA
         n.trial <- NA
        }
        if(pr){print(c(go, win, stop, n.trial))}
        return(c(go, win, stop, n.trial))
}
#### SIMULATES ONE TRIAL
sim.trial <- function(design.parameters, true.hist.pos, true.sens,
true.spec, pr=F){

     ### simulation data
     data <- make.data(max(design.parameters$look), true.hist.pos,
true.sens, true.spec)
     go <- 1
     look <- 1
     while(go==1){
       n.at.look <- design.parameters$look[look]
       test.statistics <- primary.analysis(n.at.look, data, design.
parameters,pr)
       trial.result <- stop.check(test.statistics, design.parameters,pr)
       go <- trial.result[1]
       if(go==1){
                     look <- look + 1
       }
     }
     # Return
     # 1. Win=1, Lose=0
```

```
     # 2. Why stop, 1=Futility, 2=Max N, 3=Early Success
     # 3. Sample size
     # 4. Observed sensitivity
     # 5. Observed specificity
     # 6. Post Prob Sens meets goal
     # 7. Post Prob Spec meets goal

     N <- trial.result[4]
     return(c(trial.result[2:4],      mean(data[1:N,2][data[1:N,1]==1]),
                                      1-mean(data[1:N,2][data[1:N,1]==0]),
                                      test.statistics[3:4]))
}
##### SIMULATES Nsims trials and summarizes results.
##### if long=c(T,T,T,T,T) then it produces more output.
##### turn each to F to turn off.
##### will print details for first "pr" simulated trials
sim.N.trials.summarize <- function(Nsims, design.parameters, true.hist.pos,
     true.sens, true.spec, pr=10, long=rep(T,5)){
  results.matrix <- matrix(nrow=Nsims, ncol=7)
  for(s in 1:Nsims){
  results.matrix[s,] <- sim.trial(design.parameters, true.hist.pos,
true.sens, true.spec, s<=pr)
     }
  t1 <- round(table(factor(results.matrix[,1], levels=0:1,
     labels=c("Fail","Success")))/Nsims, 3)
  t2 <- round(table(factor(results.matrix[,2], levels=1:3,
     labels=c("Futility","MaxN","EarlySuccess")))/Nsims, 3)
  t3 <- round(table(factor(results.matrix[,3],design.parameters$looks),
     factor(results.matrix[,1],levels=0:1, labels=c("Fail", "Success")))/
     Nsims, 3)
  t4 <- round(table(factor(results.matrix[,3],design.parameters$looks),
     factor(results.matrix[,2],levels=1:3,
     labels=c("Futility","MaxN","EarlySuccess")))/Nsims, 3)
  t5 <- round(table(
     factor(as.numeric(results.matrix[,6] >= design.parameters$sens.critv),
levels=0:1,
     labels=c("Lose Sens","Win Sens")),
     factor(as.numeric(results.matrix[,7] >= design.parameters$spec.critv),
levels=0:1,
     labels=c("Lose Spec","Win Spec")))/Nsims, 3)

     if(long[1]){print(t1)}
     if(long[2]){print(t2)}
     if(long[3]){print(t3)}
     if(long[4]){print(t4)}
     if(long[5]){print(t5)}
     cat("\n\n")
```

```
 return(c(Power=mean(results.matrix[,1]), PrFutStop=mean(results.matrix
 [,2]==1)
, PrMaxN=mean(results.matrix[,2]==2)
, PrEarlySuc=mean(results.matrix[,2]==3)
, MeanN=mean(results.matrix[,3])
, MeanSens=mean(results.matrix[,4])
, MeanSpec=mean(results.matrix[,5])
))
     }
```

REFERENCES

AtriCure. (n.d.). AtriCure bipolar radiofrequency ablation of permanent atrial fibrillation (ABLATE). http://clinicaltrials.gov/ct2/show/NCT00560885 (accessed November 20, 2012).

Berry, D. A. (1989). Monitoring accumulating data in a clinical trial. *Biometrics* 45(4):1197–1211.

Berry, D. A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science* 19(1):175–187.

Choi, S. C., Pepple, P. A. (1989). Monitoring clinical trials based on predictive probability of significance. *Biometrics* 45(1):317–323.

Dmitrienko, A., Wang, M.-D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine* 25(13):2178–2195.

Food and Drug Administration. (2010). Guidance for industry adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatory Information/Guidances/ucm201790.pdf

Gaydos, B., Anderson, K. M., Berry, D., Burnham, N., Chuang-Stein, C., Dudinak, J., Fardipour, P., et al. (2009). Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal* 43(5):539–556.

Gelman, A, Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Hampson, L. V. and Jennison, C. (2013). Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society, Series B* 75(1):3–54.

He, W., Kuznetsova, O. M., Harmer, M., Leahy, C., Anderson, K., Dossin, N., Li, L., Bolognese, J., Tymofyeyev, Y., Schindler, J. (2012). Practical considerations and strategies for executing adaptive clinical trials. *Drug Information Journal* 46(2):160–174.

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

Julian, T. B., Blumencranz, P., Deck, K., Whitworth, P., Berry, D. A., Berry, S. M., Rosenberg, A., et al. (2008). Novel intraoperative molecular test for sentinel lymph node metastases in patients with early-stage breast cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 26(20):3338–3345.

Kelly, C. M., Green, M. C., Broglio, K., Thomas, E. S., Brewster, A. M., Valero, V., Ibrahim, N. K., et al. (2012). Phase III trial evaluating weekly paclitaxel versus docetaxel in combination with capecitabine in operable breast cancer. *Journal of Clinical Oncology* 30(9): 930–935.

Muss, H. B., Berry, D. A., Cirrincione, C. T., Theodoulou, M., Mauer, A. M., Kornblith, A. B., Partridge, A. H., et al. (2009). Adjuvant chemotherapy in older women with early-stage breast cancer. *New England Journal of Medicine* 360(20):2055–2065.

Oncolytics Biotech. (n.d.). Efficacy study of REOLYSIN® in combination with paclitaxel and carboplatin in platinum-refractory head and neck cancers. http://clinicaltrials.gov/ct2/show/NCT00560885 (accessed November 19, 2012).

Patient-Centered Outcomes Research Institute. (2012). *Draft Methodology Report: "Our Questions, Our Decisions: Standards for Patient-centered Outcomes Research."* Draft. Patient-Centered Outcomes Research Institute. http://pcori.org/assets/MethodologyReport-Comment.pdf (accessed September 20, 2012).

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Saville, B. R., Connor, J. T., Ayers, G. D., Alvarez, J. (in press). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*.

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

White, W. B., Grady, D., Giudice, L. C., Berry, S. M., Zborowski, J., Snabes, M. C. (2012). A cardiovascular safety study of LibiGel (testosterone gel) in postmenopausal women with elevated cardiovascular risk and hypoactive sexual desire disorder. *American Heart Journal* 163(1):27–32.

Wilber, D. J., Pappone, C., Neuzil, P., De Paola, A., Marchlinski, F., Natale, A., Macle, L., Daoud, E. G., Calkins, H., Hall, B., Reddy, V., Augello, G., Reynolds, M. R., Vinekar, C., Liu, C. Y., Berry, S. M., Berry, D. A., for the ThermoCool AF Trial Investigators. (2010). Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation: A randomized controlled trial. *Journal of the American Medical Association* 303(4):333–340.