Frequentist Operating Characteristics of Bayesian Posterior Probability Designs for Medical Device Trials that Include a Single, Late-information-time, Interim Analysis

Greg Maislin[1]
[1]Biomedical Statistical Consulting, 1357 Garden Road, Wynnewood, PA 19096

Abstract

In medical device trials, a single, late-information-time interim analysis can substantially reduce time to regulatory filing. Moreover, candidate control devices are often the subject of their own recent regulatory study, if not of multiple studies. Entirely ignoring this information at the time of analysis is increasingly difficult to justify. In this paper, Bayesian posterior probability designs are proposed that address these issues. Design goals include retention of some randomization to address hidden bias while allowing a substantial reduction in the control sample size through the use of a properly calibrated informative Bayesian prior. The FDA Guidance on Bayesian trials (February 2010) stipulates the need to explore frequentist operating characteristics of such designs. The numbers of draws needed to achieve acceptably low rates of simulation error will be evaluated for the case of non-inferiority trials utilizing a composite clinical success endpoint. Characteristics of designs with and without using a control informative prior will be compared and assessed as functions of sample size and how informative the control prior is assumed to be.

Key Words: Bayesian design, posterior probability, interim analysis, frequentist operating characteristics, informative prior, orthopaedic device

1. Introduction

There is great need for creative thinking in the design of medical device trials to improve efficiency while maintaining scientific integrity of clinical trial results. Bayesian approaches to clinical trial design offer additional flexibility with regard to design features but there remains concern about the validity of results within the regulatory setting. FDA Guidance on Bayesian trials (February 2010) indicates that demonstration of Bayesian test operating characteristics from a frequentist perspective (i.e., expected type I and type II error rates) is an essential element in the evaluation of the validity of Bayesian inference for regulatory studies. This paper discusses a class of randomized clinical trial designs for demonstration of clinical non-inferiority between an investigational device (I) and a control device (C) based on a composite clinical success (CCS) criterion evaluated post operatively. For simplicity, examples in this paper assume clinical status will be evaluated at Month 24 post operatively. The trial design includes a single, late-information-time interim analysis (IA); the purpose of which is to substantially reduce time to regulatory filing. Moreover, candidate control devices are often the subject of their own recent regulatory study, if not of multiple studies. Entirely ignoring this information at the time of analysis is increasingly difficult to justify when resources are not unlimited. To accommodate the use of this information, the Bayesian

posterior probability design is extended to incorporate an informative prior on the control device only. Design goals include retention of some randomization to address hidden bias while allowing a substantial reduction in the control sample size through the use of a properly calibrated Bayesian informative prior. Frequentist operating characteristics of designs with and without using a control informative (c-informative) prior for several example designs are explored and compared as functions of sample size and how informative the control prior is assumed to be. Implementation of the simulations is discussed using SAS and R. The number of draws needed to achieve acceptably low rates of simulation error is also evaluated.

2. Modelling Considerations

2.1 Candidate approaches

There are three approaches to Bayesian interim analysis (IA) described in the FDA guidance (2010). One approach focuses on decision analyses that consider the "cost of decision errors and experimentation in deciding whether to stop early". This approach is not considered here. Another approach is the predictive probability approach in which Bayesian computations are used to determine the distribution of the 'as yet unobserved' clinical endpoints at the time of the IA. This distribution is used in estimating the (predictive) probability that the original frequentist hypothesis testing criteria will or will not be met at the final analysis. For example, the predictive probability of interest at the time of IA might be the probability the lower bound of a 95% one-sided (non-inferiority) confidence interval exceeds the pre specified non-inferiority delta (e.g., δ=-0.10). If this probability is extremely high at the time of IA then the final results are essentially a foregone conclusion and so, it is argued, there is sufficient evidence for supporting the efficacy claim on the basis of the IA alone. In this sense, the predictive probability approach is a hybrid approach sharing both frequentist and Bayesian aspects. In contrast, a Bayesian posterior probability approach requires evaluation of the probability that most clearly relates to the regulatory question at hand. This question is, “What is the probability that the investigational device (I) is clinically inferior to the control device (C)? More precisely, “What is the probability that the Month 24 CCS rate for I is more than δ smaller than C. The definition of posterior probability requires no appeal to frequentist decision rules for its evaluation and so in this sense is “more Bayesian” than the predictive probability approach. However, in both cases, once the rule is defined, simulations must be performed to determine their respective frequentist operation characteristics (type I and type II error rates).

2.1 Rationale for single, late-information-time IA

The rationale for performing an IA is to reduce Sponsor burden associated with submission time delay that results from having waiting for all efficacy evaluable patients to become theoretically due for Month 24 follow-up. Design parameters may be tuned such that the experimental design has a relatively high probability of permitting early submission (i.e., reaching a conclusion of clinical non-inferiority at IA), yet controls type 1 (α) error at an acceptable level as demonstrated through simulation.

There is another important issue that motivates the use of a late-information-time IA. Maintaining double-blind in randomized clinical trials of investigational medical devices is often not feasible and if feasible is not practical for technical reasons. Ideally, an IA is performed and results interpreted by an ‘independent third party’ in order to not influence subsequent behaviour by patients or investigators. In order to eliminate the possibility that dissemination of IA results influence future enrolment patterns, the IA should not be

conducted until the targeted maximum sample sizes in each device group have been randomized. In this way it is not possible for release of IA results to influence future study enrolment, thus avoiding selection bias. For example, if the last patient is randomized 6 months before the planned IA, then more than 18 months would have to pass before the opportunity arises for final analysis. If there is sufficient and substantial evidence of non-inferiority at the time of the IA, then not performing the IA represents a lost opportunity to substantially shrink the timeline necessary to bring safe and effective medical devices to those who need them. The benefit to the Sponsor is not really in terms total study cost. This is because the study design requires completing follow-up as appropriate for all randomized patients and re-computing the Bayesian posterior probability after all patients are theoretically due for Month 24 follow-up. All of the examples in this paper assume that the IA will take place when 70% of the target sample size for the efficacy evaluable cohort is achieved. Depending upon enrolment rates, this percentage might need changing to insure that the IA occurs after the last patient is randomized.

2.2 Rationale for use of an informative control prior

It is natural to use a non-informative prior distribution on Month 24 CCS for I. In contrast, candidate C devices are often the subject of their own recent regulatory study, if not of multiple studies. Increasingly, there may be candidate C's for which sufficient summary data is available to guide construction of a reasonable priori. Entirely ignoring this information at the time of analysis is increasingly difficult to justify when resources are not unlimited. The Bayesian posterior probability design with a single late-information-time IA is proposed to address this issue. Design goals include retention of some randomization to address hidden bias while allowing a substantial reduction in the control sample size through the use of a properly calibrated informative Bayesian prior.

In this way, the Bayesian approach improves efficiency by not ignoring valid scientific evidence concerning control device performance when making comparisons to a new investigational device. The Bayesian approach includes a formal way to incorporate such information in treatment group comparisons and influence statistical findings regarding clinical non-inferiority. This is in distinction to the frequentist approach which only permits use of such information at the study design phase but not at the analysis phase. As a consequence, it is critical for the Sponsor to sufficiently justify the particular prior distribution to be used and to reach consensus with FDA regarding this prior.

2.3 Outline of the Non-Inferiority Testing Procedure

The methods used rely on the relatively familiar and straight-forward statistical model based on beta-binomial distributions. Defining pI and pC as the probability of achieving Month 24 CCS for I and C, respectively, then prior beta prior distributions on pI and pC will be updated by the observed data (XI,NI) and (XC,NC) to produce the posterior distributions for pI and pC; and hence the posterior distribution for the device group difference in probabilities of Month 24 CCS. Here XI and XC refer to the numbers of patients that achieve Month 24 CCS in I and C, respectively. Analogously, NI and NC refer to the number of patients evaluated in the two groups. The prior on I is assumed to be the non-informative Jeffries beta(0.5,0.5) distribution. This makes particular sense for an investigational device under regulatory review.

In contrast, there may be substantial information is available for C. Bayesian designs permit the formal use of this information through the specification of an informative prior distribution on pC. The use of this information reduces the number of patients that need

to be randomized to C. The impact of using a "C-informative prior" is assessed below by comparing sample sizes and frequentist operating characteristics for designs that use and do not use a C-informative prior and by changing the amount of 'discounting' prior information by varying degrees.

Because the beta distribution is the conjugate prior for the binomial distribution, the posterior distributions are also beta distributions. Parameters of 'updated' beta distributions are (a+X, b+(N-X)) where X is the number of evaluable patients achieving success and N-X is the number of evaluable patients not achieving success. The posterior distribution of the group difference can be determined from the difference in these appropriately updated beta distributions. The range of this distribution is -1 to +1. The integral from -0.10 to 1 is the Bayesian posterior probability that the Month 24 CCS rate for patients implanted with I is no more than -0.10 less than that for C.

2.4 Exchangeability

There are a number of potential approaches for objectively translating prior clinical evidence into the parameters of a specified prior distribution. A fundamental challenge to the effective use of prior studies involves 'exchangeability'. There are two levels of exchangeability, patients within trials and among trials. "Units (patients or trials) are considered exchangeable if the probability of observing any particular set of outcomes on those units is invariant to any reordering of the units." (FDA Guidance 2010)

In most instances exchangeability can be assumed within trials simply based on their design and conduct. Exchangeability among trials is most easily represented by imagining that trials arise from some 'super population' and that expected results are invariant to any reordering but may randomly vary among studies around some central tendency. The amount of prior information available for use can be formally related to parameters of this "super population" using random effects modeling. In such models, the amount of information available for use in the new study is inversely proportional to the variance among trials. It is intuitively clear that one should place more trust (and weight) on prior results when multiple studies produce very similar results and should be suspicious (and place less weight) on wildly varying prior results.

The random effects approach is appealing and might be a candidate as a gold standard benchmark that other proposed methods can be compared against. However, multiple studies are typically not available for medical devices while at the same time substantial information concerning the control device might be available through its own FDA approval package. Such studies may be expected to be similar to the study being designed on the basis of its restricted indication, similar study conduct, and similar types of institutions and investigators.

Although the above reasoning suggests at least reasonably good exchangeability, there is still need to discount the prior information to account for lack of perfect exchangeability; and also to potentially reduce to impact of the prior on the Bayesian posterior distribution to more modest levels (i.e., to 'calibrate' the prior). This calibration is necessary to avoiding priors that are "too informative" yet allow Sponsors a means to appropriately utilize prior information in both the design and analysis of trial data. This in turn provides the opportunity to reduce Sponsor burden (in terms of number of total patients enrolled and therefore, total cost) when providing the evidence necessary for regulatory evaluation device effectiveness as mandated by Section 513(a)(3) of the Federal Food, Drug, and Cosmetic Act (FFDCA) (see 21 U.S.C. 360c(a)(3)).

4. Implementation of Design

4.1 Statistical model

The approach adopted here was formulated in terms of determining the posterior probability of non-inferiority at the time of IA and then again at the final analysis if there was insufficient evidence favouring non-inferiority at the time of IA. The following steps assume non-informative priors for both devices. A subsequent section examines the impact of an informative prior for the control device only. The description of following algorithm assumes that the criterion for rejecting the inferiority hypothesis at the IA is a posterior probability of non-inferiority that is at least equal to 0.95. If the posterior probability is less than 0.95, it will be recomputed after adding in data from the remaining efficacy evaluable patients after they all become theoretically due for Month 24 follow-up. The actual Bayesian posterior thresholds used at IA and final analysis should be thought of as tuning parameters that allow designs to vary in terms of type I and type II error.

1. Assume non-informative (Jeffries) prior distributions for the success likelihoods of the investigational (pI) and control device (pC) groups.
2. Using results from an interim analysis, determine the Bayesian posterior bivariate distribution for the difference (pI-pC). Individually, pI and pC have beta distributions. The needed bivariate distribution can be empirically obtained through simulations in which a value for pI is obtained from its (device group specific) distribution and a value for pC is obtained from its distribution. The difference (pI-pC) may then be computed. If this process is repeated a large number of times, the resulting frequency distribution of observed (pC-pI) is an empirical estimate of the needed probability distribution.
3. From the bivariate posterior distribution obtained in step 2, determine the probability that $(pI-pC) > -\delta$, i.e., that is $(pI-pC) > -0.10$. This probability can be determined by counting the number of simulations in which this condition is met.
4. If the 5th percentile of the bivariate distribution of (pI-pC) exceeds -0.10, then the probability that the success rate among I is no more than 0.10 less than C is at least 0.95. This finding supports the decision to allow early device registration and approval since sufficient information is available at the time of the IA to support the non-inferiority claim.
5. If the 5th percentile of (pI-pC) does not exceed -0.10, then the probability that the success rate for I is no more than 0.10 less than for C is smaller than 0.95. Therefore, the endpoint status for the remaining evaluable patients must be obtained and a final posterior probability computation performed after all such patients are theoretically due for Month 24 follow-up by repeating the steps described above.

4.2 Software implementation - SAS

The first example is a study designed to demonstrate clinical non-inferiority of a new investigational medical device relative to control. A 2:1 blocked randomization was planned in order to have more safety data on the investigational device. At the design stage it was assumed under non-inferiority that both devices had a Month 24 CCS rate equal to 0.85. The Sponsor expected that the investigational device would be at least slightly better. This is an important consideration because as shown later, even a true slight superiority of I relative C increases the chance for early registration substantially.

Therefore, three scenarios were initially evaluated. These were (pI0, pC0) = (0.85, 0.85), (0.74, 0.85), and (0.87, 0.85). These reflect true equivalence, true inferiority, and slight superiority. The first and third scenarios were used to obtain estimates of statistical power (i.e., 1 minus type II error) and the second was used to obtain estimates of type I error. Subsequently, the case of (ps0, pc0) = (0.75, 0.85), or "borderline" inferiority was also evaluated. This case was compared with one in which the IA posterior probability criterion was increased from 0.95 to 0.975 as an example of 'tuning' design parameters to improve frequentist operating characteristics.

The algorithm described above was implemented first in SAS and then in R. The logic of the SAS algorithm is sequential processing and so more easily explained in text. The sequential logic is outlined below. R is based on vector arithmetic. The code used to generate the simulations is provided. Obtaining the same results up to simulation error provided evidence of programming validity.

1. Choose a study design based on NI1, NI2, NC1, and NC2: NI1 = planned sample size for I at the time if IA; NI2 = number of additional I to be evaluated during final analyses; NC1 = planned sample size for C at IA; and NC2 = number of additional C to be evaluated at final analysis.
2. Fix pI0 = true success rate for I and pC0 = true success rate for C.
3. Start 'external loop' (explained below).
4. Based on (NI1, NC1) and (pI0, pC0) obtain simulations for XI1 and XC1) at the time of IA. [e.g., XI1=rand('binomial',pI0,nI1); XC1=rand('binomial',pC0,NC1);]
5. Determine parameters for interim posterior distributions for pI1 and pC1. aI1 = (XI1 + .5); bI1 = (NI1 - XI1 + .5); aC1 = (XC1 + .5); bC1 = (NC1 - XC1 + .5);
6. Perform an 'internal loop' to obtain the posterior distribution for (pI1-pc2): do obs=1 to &numsim; pI1=rand('beta',aI1,bI1); pC1=rand('beta',aC1,bC1); postdiff1 = pI1 - pC1;
7. Use this inner loop to determine the probability distribution of the difference between two beta variables and the probability that the group difference exceeds -0.10. This step will be used in actual data analysis. In contrast, the outer loop is solely for simulations needed to determine the operating characteristics of the Bayesian interim analysis plan.
8. Obtain simulated final (post IA) posterior distributions. In applications, the posterior distribution for the final analysis would be based on binomial distributions with parameters (pI1, pC1). However, in simulations, there is need to assume exchangeability. Therefore, the simulated numbers of successes for the additional subjects added after IA were determined using the original assumed true values (pI0 and pC0) and not those implied by Bayesian updating: XI2=rand('binomial',pI0,NI2); XC2=rand('binomial',pC0,NC2). The parameters for the final posterior distributions were then simulated as: aI2=aI1 + XI2; bI2=bI1 + (NI2 - XI2); aC2=aC1 + XC2; bC2=bC1 + (NC2 - XC2); The final posterior distribution for (pI2-pC2) was similarly determined as described above: pI2=rand('beta',aI2,bI2); pC2=rand('beta',aC2,bC2); postdiff2 = pI2 - pC2;
9. Iterate the 'external loop': Given (pI0, pC0), simulate values for (XI1, XC1|NI1, NC1) and separately for (XI2, XC2|NI2, NC2) multiple times. For each of these times, the internal loop provided the necessary percentiles for the distributions of (pI1-pC1) and (pI2-pC2).

10. Tabulate results: Determine which of the following cases occurred for each iteration of the external loop:

a) The hypothesis of non-inferiority is rejected at the interim analysis.

b) Among cases in which the non-inferiority hypothesis is not rejected at the interim analysis, it is rejected at the final analysis.

c) The non-inferiority hypothesis is not rejected at the interim analysis or at the final analysis.

The following output is from a typical scenario. For this example, it was assumed that there was true inferiority defined by pI0=0.74 and pC0=0.85. Sample sizes of 140 and 70 were chosen for IA. Data from the final 60 and 30 efficacy evaluable patients (already randomized) will be accumulated and if a conclusion of non-inferiority cannot be reached at IA, the same Bayesian posterior probability will be computed based on N=200 and N=100 I and C, respectively. Both inner and outer loops were set to 3000 in this simulation.

Bayes_interim_sim:
pI0=0.74, pC0=0.85, NI1=140, NC1=70, NI2=60, NC2=30
numsim=3000, delta=-0.10, SIMULATED SAMPLES NUMBER = 3000
Summary of Interim/Final Credible Interval Simulations

Cumulative REJECTCODE_1SIDED	Frequency	Percent	Cumulative Frequency	Percent
Reject at interim	102	3.40	102	3.40
Reject at final	52	1.73	154	5.13
Fail to reject	2846	94.87	3000	100.00

This study design maintains close to a type I error rate of 0.05 (5.13%) when pI0=0.74 and pC0=0.85. If the true success probabilities are both actually equal to 0.85, then the estimated probability of achieving the early registration criterion is 0.664 as shown in the next exhibit. There is a 0.156 probability that the study will fail to reject non-inferiority during IA, but then be able to reject non-inferiority during the final analysis. Therefore, the overall statistical power is estimated to be 82.0%.

Bayes_interim_sim:
pI0=0.85, pC0=0.85, NI1=140, NC1=70, NI2=60, NC2=30
numsim=3000, delta=-0.10, SIMULATED SAMPLES NUMBER = 3000
Summary of Interim/Final Credible Interval Simulations

Cumulative REJECTCODE_1SIDED	Frequency	Percent	Cumulative Frequency	Percent
Reject at interim	1991	66.37	1991	66.37
Reject at final	468	15.60	2459	81.97
Fail to reject	541	18.03	3000	100.00

However, if I is, in fact, slightly better than C (i.e., pI0=0.87 and pC0=0.85), the probability of early stopping increases from 0.663 to 0.759. Additionally, there is a 0.146 probability that the study will fail to reject non-inferiority at IA, but then be able to reject non-inferiority during the final analysis. Therefore, the overall statistical power for this scenario is 90.5%.

Bayes_interim_sim: pI0=0.87, pC0=0.85, NI1=140, nc1=60, ns2=60, nc2=30
numsim=3000, delta=-0.10, SIMULATED SAMPLES NUMBER = 3000
Summary of Interim/Final Credible Interval Simulations

REJECTCODE_1SIDED	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Reject at interim	2278	75.93	2278	75.93
Reject at final	438	14.60	2716	90.53
Fail to reject	284	9.47	3000	100.00

Results for 'borderline inferiority' defined as pI0=0.75, pC0=0.85 were:

Bayes_interim_sim: pI0=0.75, pC0=0.85, NI1=140, NC1=70, NI2=60, NC2=30
numsim=3000, delta=-0.10, SIMULATED SAMPLES NUMBER = 3000
Summary of Interim/Final Credible Interval Simulations

REJECTCODE_1SIDED	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Reject at interim	154	5.13	154	5.13
Reject at final	65	2.17	219	7.30
Fail to reject	2781	92.70	3000	100.00

For this scenario, type I error is a little larger than 5% (7.3%). Type I error may be controlled by adjusting the posterior probability thresholds. For example, the IA threshold could be increased from 0.95 to 0.975 as illustrated below.

4.3 Software implementation - R

In order to have more flexibility with regard to tuning parameters, the algorithm was reprogrammed using R and this code is provided below.

```
library(lattice)
posterior <-function(ai0,bi0,ac0,bc0,xi1,ni1,xc1,nc1,delta,numsim) {
ai1 <- ai0 + xi1
bi1 <- bi0 + (ni1 - xi1)
ac1 <- ac0 + xc1
bc1 <- bc0 + (nc1 - xc1)

pi1 <- rbeta(numsim,ai1,bi1)
pc1 <- rbeta(numsim,ac1,bc1)

pstdif <- pi1 - pc1 ;
pstdif.ge.delta <- pstdif > delta
sum.pstdif.ge.delta <- sum(pstdif.ge.delta)

posterior.prob <- sum.pstdif.ge.delta / length(pstdif.ge.delta)

rate_i <- xi1/ni1
rate_c <- xc1/nc1
rates  <- c(rate_i, rate_c)

require(stats)

out=list(Rates=rates,
       Posterior_Prob=posterior.prob,
       histogram(pstdif, nint=50,xlab='Delta',main="Posterior Distribution for
Device Difference in Mo. 24 CCS")
)
out
}
gm_adapt2(hypPs=0.85,hypPc=0.85,simPs=0.85,simPc=0.85,ns1=140,nc1=70,ns2=60,nc2=30
,as0=0.5,bs0=0.5,ac0=0.5,bc0=0.5,delta<-0.10,postc1<-.95,postc2<-.95)

Binomial Draws  Beta Draw  IA Post.>  Final Post.>     Set Ps
"5000.000"      5000.000"  "0.950"     "0.950"        "0.850"

Set Pc   IA Stopping Prob Cond. Final Reject. Prob  Overall Reject. Prob   Max N
"0.850"    "0.6566"              "0.148"                "0.8046"       "300.000"

Expected N
"240.906"
```

The results for IA, final, and cumulative power based on 5000/5000 beta/binomial draws in R are very similar to the results from the SAS implementation based on 3000/3000 beta/binomial draws. The simulation results for type I error for the case of pI=0.75 and pC=0.85 is provided below. Results are again similar to those obtained using SAS.

```
gm_adapt2(hypPs=0.85,hypPc=0.85,simPs=0.75,simPc=0.85,ns1=140,nc1=70,ns2=60,nc2=30
,as0=0.5,bs0=0.5,ac0=0.5,bc0=0.5,delta<-0.10,postc1<-.95,postc2<-.95)

Binomial Draws  Beta Draw    IA Post.>  Final Post.>     Set Ps
 "5000.000"     "5000.000"   "0.950"    "0.950"          "0.750"

Set Pc  IA Stopping Prob Cond. Final Reject. Prob  Overall Reject. Prob  Max N
 "0.850"       "0.049"              "0.019"              "0.068"    "300.000"

Expected N
"295.590"
```

Expected type I error under these assumptions can be reduced from 0.068 to 0.062 by increasing the IA posterior probability criterion from 0.95 to 0.975.

```
gm_adapt2(hypPs=0.85,hypPc=0.85,simPs=0.75,simPc=0.85,ns1=140,nc1=70,ns2=60,nc2=30
,as0=0.5,bs0=0.5,ac0=0.5,bc0=0.5,delta<-0.10,postc1<-.975,postc2<-.95)

Binomial Draws   Beta Draw   IA Post.>  Final Post.>   Set Ps
"5000.000"       "5000.000"  "0.975"    "0.950"        "0.750"

Set Pc   IA Stopping Prob Cond. Final Reject. Prob  Overall Reject. Prob  Max N
"0.850"      "0.0274"         "0.0352"       "0.0623"           "300.000"

Expected N
"298.390"
```

4.4 Simulation Error

Simulation error was assessed by re-running particular scenarios 10 times and computing the standard deviation across the 10 simulations of type I and type II error at the interim analysis, at the final analysis, and cumulatively.

The first scenario is summarized below:

- X beta / X conditional binomial draws per iteration with X={1000, 2000, 3000, 4000, and 5000)
- A single IA at 70% information time (the '140/70 + 60/30' design)
- Month 24 CCS rate in I is 0.85; Month 24 CCS rate in C is 0.85
- IA posterior criterion = 0.95; final analysis posterior criterion = 0.95

Table 1 summarizes the 10 simulations conducted in order to demonstrate that simulation error is adequately controlled for when using 5000 beta / 5000 binomial draws. It took approximately two days on a quad core PC running VISTA and using R 2.11.1.

Over ten independent estimates of power, the mean (SD) was 0.814 (0.0075) with minimum value 0.799 and 0.826. If the population of simulated estimates were normally with $\mu=0.814$ and $\sigma=0.0075$, then 95% of the determinations would fall from 0.80 to 0.83 to two decimal places. Thus, simulation is low for 5000/5000 iterations. This process was repeated varying the numbers of draws with results summarized in Figure A and then repeated for the case of borderline inferiority in Figure B.

Table 1
Determination of Simulation Error of Frequentist Power
for a Bayesian Posterior Probability Design
with a Single Interim Analysis at 70% Information Time
5000 betas / 5000 conditional binomials per iteration
Rate I = 0.85, Rate C = 0.85
IA Post Prob Criterion = 0.95, Final Post Prob Criterion = 0.95

Simlation	Prob(Reject at IA)	Prob(Reject at final)	Total Prob(Reject)
1	0.6620	0.1550	0.8170
2	0.6656	0.1522	0.8178
3	0.6760	0.1500	0.8260
4	0.6467	0.1518	0.7994
5	0.6588	0.1538	0.8124
6	0.6548	0.1588	0.8136
7	0.6514	0.1568	0.8082
8	0.6600	0.1510	0.8110
9	0.6110	0.1612	0.8222
10	0.6564	0.1608	0.8172
mean	0.6543	0.1551	0.8145
SD	0.0172	0.0041	0.0075
min	0.6110	0.1500	0.7994
max	0.6760	0.1612	0.8260

Figure A

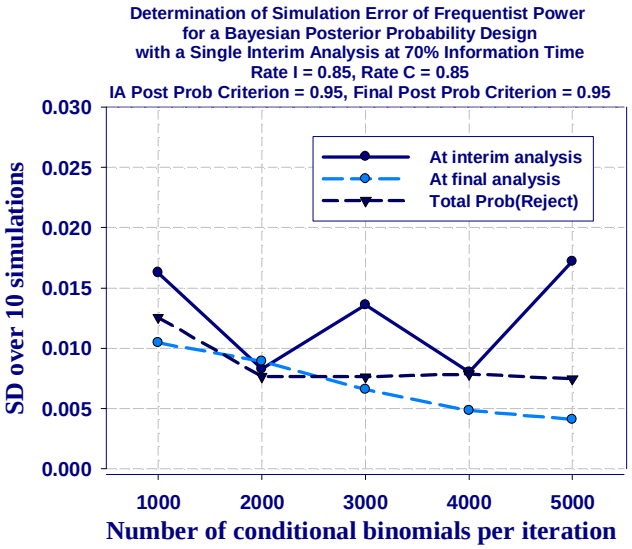


Figure B

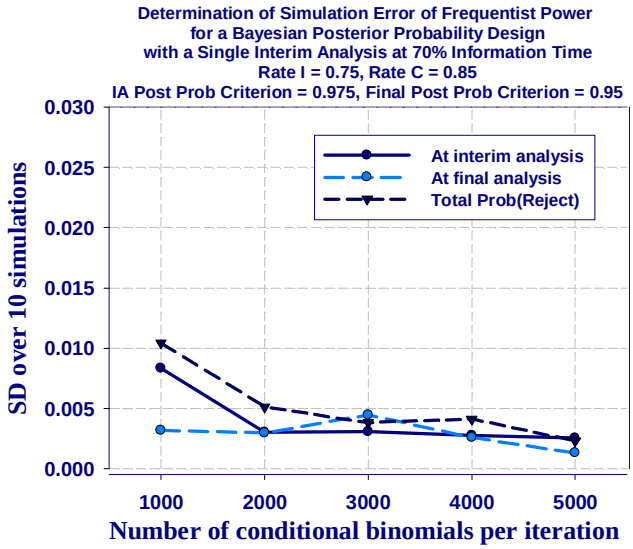


Figure A displays a nearly linear relationship between simulation error and the number of iterations when estimating (conditional) statistical power at the final analysis. The simulation error for cumulative power was substantially reduced by increasing the number of interactions from 1000 to 2000, but then did not subsequently decrease. The simulation error estimates for power at the IA did appear to decrease as a function of number of iterations for this example.

Figure B shows that simulation error for type 1 error is smaller than for power. Simulation error of IA type I error and total type I error were substantially reduced using

2000 iterations rather than 1000. There appeared to be additional smaller reductions in cumulative type 1 error for increases in the numbers of simulations from 2000 to 3000 or 4000 and then another smaller reduction going from 4000 to 5000. IA type I simulation error also dropped considerably going from 1000 to 2000 iterations, but did not further appreciably drop when the number of iterations was increased further. (Conditional) final analysis type I error appeared least affected by increasing the number of iterations.

5. Design using Informative Control Prior

5.1 Specifics of Approach

The objective was to determine reasonable values of the (a,b) parameters for the control informative prior beta distribution. The parameters (a,b) are often interpreted as the reflecting the prior number of 'successes' and 'failures' since the number of successes is added to the current value of "a" in order to obtain an updated value for a; and similarly for "b". The mean of the beta distribution is a/(a+b) which is interpreted as the prior estimate of the event probability. Using Bayes theorem, the posterior distribution can be determined by multiplying the prior distribution by the usual binominal likelihood function conditional on these beta parameter values. In this study, since the endpoint is binary, the binomial likelihood is used. Since the beta distribution is the conjugate prior of the binomial, the resulting posterior distribution is also a beta distribution but with parameters a1 = a0 + X and b1 = b0 + (N-X) where X is the number of successes in a sample size equal to N.

To illustrate the 'c-informative prior' approach, we assume that an SS&E is available for the proposed control group that includes (say) N=111 efficacy evaluable patients and among these 111 patients, 80 (72%) achieved the same Month 24 CCS criterion to be used in the new study.

The strategy used to design the c-informative prior design was to first start by designing a conventional frequentist design with 1:1 randomization with adequate statistical power. Then, the control sample size was cut in half. After that, frequentist operating characteristics for a number of c-informative prior designs were evaluated. These prior distributions all had the same expected value determined from the control SS&E but varied in the amount of information that was used by varying (a+b). The goal was to find (a+b) < 111 in order to adequately discount the control prior while maintaining acceptable frequentist operating characteristics. A grid search was performed reflecting the amount of "borrowed" information defined by beta parameter sums (a+b) equal to 0, 1, 10, 20, 30, 40, 50, 60, and 70. No design in which the sum of the beta parameters exceeded 70 was considered in order to insure there would be at least one-third discounting of the prior information. The "a" parameters of the control prior beta distributions were set to 0.72 times these values plus 0.5. The value of 0.5 comes from assuming that in the completely non-informative case, the prior distribution would be the Jeffries prior (beta(0.5, 0.5)). Similarly, the b parameters for the beta priors were set to 0.28 times these sample size values plus 0.5. Using all information from the SS&E implies a posterior beta with a=80+0.5 and with b=31+0.5. However, the largest set of beta parameters assessed was a=50.9 and b=20.1. With this interpretation, the prior analysis reflects borrowing 70 patients and so represents approximately one-third 'discount' to account for lack of perfect exchangeability, although this is only roughly the impact on information.

5.2. Operating Characteristics under Informative Control Prior

The interpretation of type 1 error is not as straight forward under informative priors compared to under non-informative priors. Because of this, it is worth noting the following two points to consider (Bayesian Guidance 2010 page 29): "For Bayesian trials, here are some points to consider regarding type I error: FDA considers type I error, along with other operating characteristics of the trial design, in evaluating submission. We strive for reasonable control of the type I error rate. An adequate characterization of the operating characteristics of any particular design may need extensive simulations. For more discussion, see Section 7.4 Technical Details.

When using prior information, it may be appropriate to control type I error at a less stringent level than when no prior information is used. For example, if the prior information if favourable, the current trial may not need to provide as much information regarding safety and effectiveness. The degree to which we might relax the type I error control is a case-by-case decision that depends on many factors, primarily the confidence we have in the prior. We may recommend discounting the historical/prior information if the prior distribution is too informative relative to the current study. What constitutes "too informative" is also a case-by-case decision".

Simulations are summarized below that provide the required characterization of the type I and type II error rates. The magnitude of the type I error is interpreted in light of the above discussion.

5.3 Calibration of Bayesian decision rule

When proposing an informative prior, FDA recommends evaluation of the prior probability of the study claim, "This is the probability of the study claim before seeing any new data, and it should not be too high. What constitutes "too high" is a case-by-case decision. In particular, we recommend the prior probability not be as high as the success criterion for the posterior probability."

FDA makes this recommendation to ensure the prior information does not overwhelm the current data, potentially creating a situation where unfavourable results from the proposed study get masked by a favourable prior distribution. In an evaluation of the prior probability of the claim, FDA will balance the informativeness of the prior against the gain in efficiency from using prior information as opposed to using non informative priors. FDA then provides guidance on approaches to 'calibrate' the prior to meet regulatory requirements. Among suggestions is: "The prior distribution of the variance can be restricted to be greater than a constant, and the constant can be varied until the prior probability of the claim is lowered to the desired value." Calibration of the prior is evaluated below.

5.4 Quantification of Control Prior Uncertainty

A prior credible interval was examined for the chosen Bayesian control-informative-prior design. This interval was computed as the 2.5th percentile to the 97.5th percentile of the prior beta distribution chosen to reflect the state of uncertainty regarding prior information about the distribution of control Month 24 composite clinical success (CCS).

A design objective was for the credible interval of this prior control distribution to be wide and viewed as reasonable and reasonably conservative. The prior control credible interval is an important quantity as it may be employed in further calibration of the prior

control distribution if stakeholders perceive that the chosen prior is too informative (or not informative enough).

5.5 Starting Point

As noted above and summarized below, simulation analyses were used to guide selection of an appropriate design and to provide sensitivity analyses regarding relevant assumptions. To provide a starting point, the following two sample sizes determinations were made using an efficient frequentist confidence interval approach. The results are based on 10,000 simulations to determine the probability that the lower bound of a one-sided 95% confidence interval for the device difference in Month 24 CCS rate will exceed the non-inferiority margin of -0.10. Computations were performed using industry standard software (nQuery Advisor 7.0 module PTE1aU-1). The results summary statements provided by this software for each of the two scenarios are provided below.
These initial frequentist sample size analyses were performed to 'anchor' the Bayesian design in terms of providing reasonable starting values.

The first design includes randomizing 260 investigational devices and 260 control devices. This is the smallest sample size necessary per group to achieve at least 80% statistical power to reject the null hypothesis that the Month 24 CCS rate for I is more than -0.10 smaller compared to C. The second design is identical except for cutting the size of the control sample by 50%.

The basic idea is to make up for the loss of power that occurs when reducing the size of the control group in half by incorporating (at an appropriate discount) what is already known about control group device performance through the use of an informative prior distribution.

With 260 subjects in the standard group and 260 subjects in the test group, the lower limit of the observed one-sided 95% confidence interval will be expected to exceed -0.10 with 81% power when the Standard proportion, pS, is 0.720 and the Test expected proportion, pT, is 0.72; results are based on 10000 simulations using the Newcombe-Wilson score method to construct the confidence interval (Newcombe RG (1988). Keeping everything else the same, the impact of reducing the control device sample size in half is to reduce statistical power from 81% to 64%. The goal of the Bayesian design is to use what is known about controls to increase this power from 64% to something larger than 80% through the use of prior information based on the approval package of the control device appropriately 'discounted' to account for imperfect exchangeability.

5.6 Bayesian Informative Priors Considered

All considered beta distributions, with the exception of the Jeffries non-informative prior, have expected values equal to a/(a+b) = ~0.72. The extra "+0.5" in the numerator and denominator of these ratios are 'shrinkage' factors reducing the rate slightly. The informative control priors vary in how much information they contain as reflected in the sum of (a+b). For the informative control priors these values are 1, 10, 20, 30, 40, 50, 60, and 70, respectively. Using the simulation strategy presented above, the following values were determined using R program exhibited above. For these simulations, 3000/3000 beta/binomial draws were used for each simulation.

Figure 1 summarizes the interim analysis frequentist power using a non-informative prior as well as and for control informative prior distributions with sums of a+b equal 1+1, 10+1, 20+1, 30+1, 40+1, 50+1, 60+1, and 70+1. The "+1" terms account for the 0.5

Figure 1
Interim Analysis Rejection Probabilities
for Different Amounts of Prior Control Information
in Terms of Beta Parameters a+b

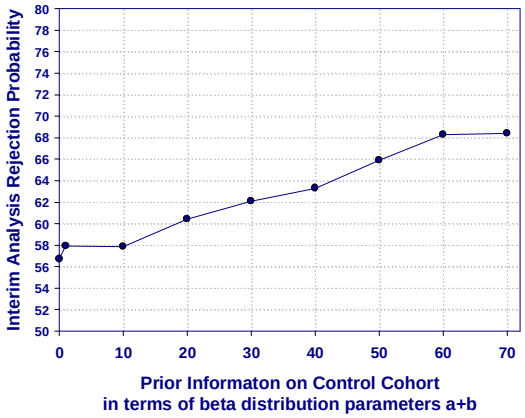


Figure 2
Cumulative Rejection Probabilities
for Different Amounts of Prior Control Information
in Terms of Beta Parameters a+b

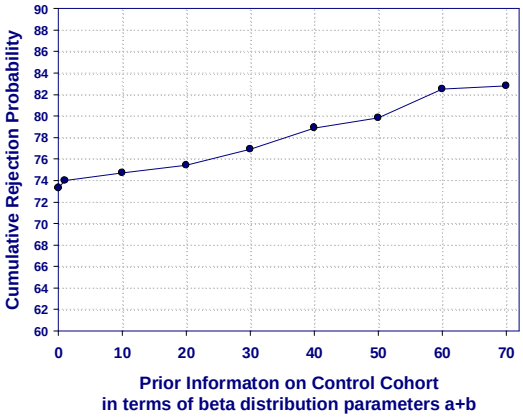
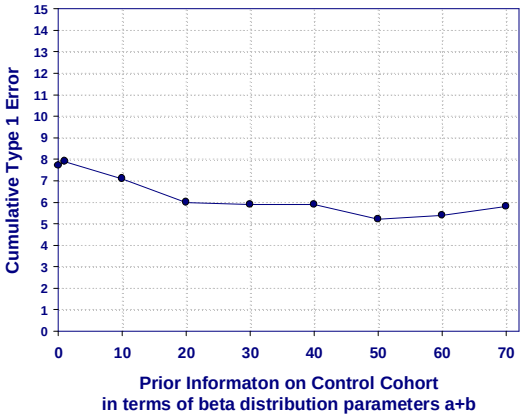


Figure 3
Cumulative Type 1 Error Rates
for Different Amounts of Prior Control Information
in Terms of Beta Parameters a+b



terms that form the Jeffries prior. IA power is increased from 56.6% to 68.3% through the use of an informative prior on the controls for the final choice of the "60+1" design.

Similarly, Figure 2 summarizes the cumulative frequentist power using a non-informative and for control informative prior distributions with sums of a+b equal 1+1, 10+1, 20+1, 30+1, 40+1, 50+1, 60+1, and 70+1. Cumulative power is increased from 73.3% to 82.7% through the use of an informative prior on the controls for the final choice of the "60+1" design.

At least 80% power is retained for a design that compares 182 I vs. 91 C at the IA on the basis of a posterior probability computed with a control informative prior defined by beta(a=43.7, b=17.3) and then includes an additional 78 I and 39 C devices to be added to the analysis sample after all evaluable patients are theoretically due for Month 24 follow-up.

Figure 3 summarizes type 1 error for each design computed assuming that the true Month 24 CCS rates were 62% and 72%, respectively, for I and C devices. For this particular design, it was not observed that there is a strong positive association between the amount of information in the prior and rate of type 1 error. For the prior defined by beta(a=43.7,b=17.3), the estimated type 1 error rate is 5.4%. Since this estimated rate is so close to 5%, it was felt that no adjustment to the Bayesian posterior probability at the time of IA had to be made in order to provide tighter control of type 1 error.

5.7 Summary of Proposed Design Characteristics

We now summarize the operating characteristics for a specific plan that ultimately would compare 260 patients randomized to I to 130 patients randomized to C. An IA would be performed when the first 182 I and 91 C are evaluable for Month 24 CCS (i.e., at 70% information time), but not before all the target enrolment of 390 patients is completed. The IA study success criterion is that the Bayesian posterior probability that Month 24 CCS rate for I is no more than 0.10 smaller than for C is at least equal to 0.95. If the IA study success criterion is not met, the Bayesian posterior probability will be recomputed after adding an additional 78 investigational and 39 control device. Assuming true equivalence (i.e., both rates are equal to 72%), then simulations demonstrate a 68.3% chance of reaching the early registration benchmark on the basis of the IA data alone with a cumulative frequentist power of 82.5%. Assuming borderline non-inferiority (Month 24 CCS equal to 62% for I and 72% for C), the frequentist type 1 error rate was estimated to be only slightly above 5% (5.4%). Simulation error was evaluated. The mean (standard deviation) across ten independent simulations of cumulative power was equal to 0.818 (0.0082). If results were normally distributed with this mean and SD, then approximately 95% of the estimated power would be between 0.802 and 0.834. Therefore, it is shown that power can be estimated sufficiently reliably based on 3000 / 3000 beta / binomial draws. Simulation error for type 1 error was again even smaller than that found for power with an SD equal to 0.0032 (mean=0.056). If results were normally distributed with this mean and SD, then approximately 95% of the estimated type 1 error rates would be between 0.048 and 0.060. Therefore, it is shown that the type 1 error rate can be estimated sufficiently reliably based on 3000 / 3000 beta / binomial draws.

5.8 Interpretation of prior distribution in terms of prior credible interval

The 97.5th percentile value of a beta (43.7,17.4) is 0.81. Similarly, the 2.5th percentile value is 0.58. Therefore, there is a 0.95 Bayesian probability that the overall success rate among C is between 0.58 and 0.81. The range from 0.58 to 0.81 reflects the prior uncertainty in the expect Month 24 CCS for C, which seems like a reasonable level of uncertainty to assume for this particular control population.

6. Conclusion

This paper shows that a Bayesian Posterior Probability design with a single late-information-time interim analysis without or with an informative control device prior has substantial potential for shortening time to approval while still providing control over type I and type II error.

Acknowledgement

The author gratefully acknowledges the support of Paradigm Spine, LLC, without which this research effort would have not been possible.

References

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, Document issued on: February 5, 2010, http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm

Newcombe RG (1988) Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine 17:873-890.