From statistical power to statistical assurance: It's time for a paradigm change in clinical trial design

Ding-Geng (Din) Chen [a,b,c] and Shuyen Ho [d]

[a]School of Social Work, University of North Carolina at Chapel Hill, NC, USA; [b]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC, USA; [c]Department of Statistics, University of Pretoria, Pretoria, South Africa; [d]PAREXEL, Durham, NC, USA

ABSTRACT

A well-designed clinical trial requires an appropriate sample size with adequate statistical power to address trial objectives. The statistical power is traditionally defined as the probability of rejecting the null hypothesis with a pre-specified true clinical treatment effect. This power is a conditional probability conditioned on the true but actually unknown effect. In practice, however, this true effect is never a fixed value. Thus, we discuss a newly proposed alternative to this conventional statistical power: *statistical assurance*, defined as the unconditional probability of rejecting the null hypothesis. This kind of assurance can then be obtained as an expected power where the expectation is based on the prior probability distribution of the unknown treatment effect, which leads to the Bayesian paradigm. In this article, we outline the transition from conventional statistical power to the newly developed assurance and discuss the computations of assurance using Monte Carlo simulation-based approach.

ARTICLE HISTORY

Received 22 November 2015
Accepted 1 November 2016

KEYWORDS

Assurance; Bayesian prior distribution; Conditional and unconditional probability; Monte Carlo simulation; Sample size determination; Statistical power

MATHEMATICS SUBJECT CLASSIFICATION

62F15; 62F03

1. Introduction

Clinical trials should be well designed for ethical consideration as well as cost effectiveness. An aspect of good design of clinical trial protocol is to determine appropriate number of patients (i.e., sample size) with adequate statistical power to address the clinical objectives. The statistical power is traditionally defined as the probability of rejecting the null if the true clinical trial treatment effect equals a prerequisite value. Therefore, the statistical power is a conditional probability to this unknown prerequisite value, as discussed extensively in Chow et al. (2003), Chen and Peace (2011), and Walter and Chen (2014). In practice, this prerequisite value is obtained from previous trials or specified based on prior experience and knowledge, which could very well be different from the true treatment effect and then could lead to an imprecise statistical power and its associated sample size.

As a result, the traditional statistical power used to "power" a clinical trial cannot actually assure a successful clinical trial. To assure a successful clinical trial, a newly proposed alternative to this conventional statistical power is "assurance," which is defined as the unconditional probability of rejecting the null hypothesis as propagated in O'Hagan and Stevens (2001), O'Hagan et al. (2005), Chuang-Stein (2006), Chow and Chang (2012), and Ren and

CONTACT Ding-Geng (Din) Chen dinchen@email.unc.edu Department of Statistics, University of Pretoria, Pretoria, South Africa.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

Oakley (2014). This assurance can then be obtained as the expected power with respect to the prior probability distribution of the prerequisite value, which leads to the Bayesian paradigm. O'Hagan et al. (2005) discussed the computational aspects with WinBUGS and gave the analytical formula for normally distributed data when the variance is known and the *Bayesian clinical trial simulation* (BCTS) when the variance is unknown. They also discussed how to apply BCTS to binary data. Chuang-Stein (2006) also illustrated the calculations for normal data with known variance using SAS/R by numerical integration. For time-to-event data, Ren and Oakley (2014) reviewed various methods and the associated calculations.

Assurance, as an alternative to the important concept of statistical power, is still new to many biostatisticians, clinicians, and government regulators. Further illustrations of this concept along with software implementation for public use remain an unmet practical need, which leads to this article. In this article, we outline the concept of assurance and discuss the computations of assurance using Monte Carlo simulation-based approach.

In Section 2, we outline the concept of transitioning from the conventional statistical power to assurance. In Section 3, we demonstrate the implementation of Monte Carlo simulation-based approach to calculate statistical power and assurance. Finally, in Section 4, a discussion is provided.

2. Conventional statistical power to assurance

2.1. Conventional statistical power and its limitations

Typically, the general objective of a clinical trial is to compare whether a new drug is better than placebo. In order to demonstrate the new drug is effective, one needs to determine how many patients should be enrolled in each treatment. In statistical terms, the null hypothesis H_0 is defined as the two treatments being not different versus the alternative hypothesis H_a is defined as the new drug being better than placebo. The hypothesis testing is then to test whether there is a statistically significant treatment effect between these two treatments. The associated concepts for this hypothesis testing are the Type I error and Type II error. The Type I error (α, typically controlled at 5%) is defined as the probability of rejecting the null hypothesis when it is true and the Type II error (β) as the probability of not rejecting the null hypothesis when it is false. The statistical power (π) is then defined as the probability of rejecting the null hypothesis when it is false (i.e., $\pi = 1 - \beta$), which is typically set between 0.8 and 0.9. The associated sample size can then be determined based on this power and the Type I error rate.

Following the notations from O'Hagan et al. (2005), we denote R as the event of rejecting the null hypothesis. The conventional definition of statistical power is then

$$\pi(\theta) = P(R|\theta), \tag{1}$$

where $\pi(.)$ is the power function and θ is a vector of the assumed parameters, such as the treatment effect, sampling variance, and possible others. It can be seen that the statistical power defined in Eq. (1) is a *conditional* probability of R conditioned on the unknown parameter vector θ. The value of this power as well as the associated sample size calculation is then dependent on the unknown parameter vector θ.

Generally, this parameter vector θ cannot be provided precisely in practical clinical trials as pointed out in O'Hagan et al. (2005) and others. Therefore, the statistical power, as one of the most important concepts in clinical trials, traditionally has been quoted as a fixed probability based on a prerequisite parameter value from the unknown alternative hypothesis parameter space. It is rare that the observed data will coincide with the prerequisite parameter value, which often lead to the issue of over-powering or under-powering a clinical trial.

2.2. Assurance in clinical trials

To eliminate these limitations from the conventional statistical power, O'Hagan and Stevens (2001) advocated the "assurance" (denoted by γ) as an alternative to this statistical power, which is defined as an *unconditional* probability to reject the null hypothesis, that is, $\gamma = P(R)$, where R is rejection of the null hypothesis. The assurance can then be viewed as the expected power to the parameter vector space of θ. It can be seen that

$$\gamma = P(R) = \int P(R,\ \theta)\, d\theta = \int P(R|\theta)\, P(\theta)\, d\theta = E_\theta\ (\mathrm{P}(\mathrm{R}|\theta))\ , \tag{2}$$

where the expectation is to the (prior) probability distribution of parameter vector space of θ.

With this definition, the "assurance" provides a bridge between the frequentists' approach in statistical power and the Bayesian paradigm of averaging or integrating out the conditional statistical power with all possible (prior) values of parameter vector space of θ. This assurance can then provide an unconditional probability or evidence to assess the success of a clinical trial and therefore is more realistic and robust than that of the conventional statistical power.

As pointed out in O'Hagan et al. (2005), the concept of assurance can be dated back to the 1980s by Spiegelhalter and Freedman (1986) and later named as a "hybrid classical-Bayesian" approach in Spiegelhalter et al. (2004). To our experience and knowledge in clinical trials, it is very reasonable to use this hybrid frequentist-Bayesian approach in study design since prior information has always been used to calculate sample size. Whenever this prior information for the unknown parameter θ (i.e., treatment effect) is sufficiently strong such that the prior variance would approach to zero, the assurance defined in Eq. (2) would approach the conventional statistical power defined in Eq. (1). On the other hand, if the prior information is weak, the prior variance would be large and the assurance defined in Eq. (1), which averages all the potential values of this vague prior distribution, would be more appropriate than the conventional statistical power to assess the probability of a successful trial.

2.3. Illustrations

Conceptually, assurance defined in Eq. (2) is the expected power to the parameter vector space of θ. Depending on the dimension of this parameter vector space, the expected power can be high-dimensional integration, which makes analytical formula virtually impossible. As an illustration from the computational aspect, we use the simple case of normally distributed data for two treatments to illustrate the transition process from statistical power to assurance.

Suppose that in a two-treatment clinical trial with n_i patients randomized to treatment i ($i = 1,\ 2$), the continuous outcome x_{ij} from jth patient is normally distributed as $x_{ij} \sim N(\mu_i,\ \sigma_i^2)$. Assuming that σ_i^2 are known and we estimate the population means with the sample means as $\bar{x}_i \sim N(\mu_i, \sigma_i^2/n_i)$. Then the treatment difference $\delta = \mu_1 - \mu_2$ can be estimated by $\hat{\delta} = \bar{x}_1 - \bar{x}_2$ and the standard deviation can be calculated as $\tau = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, that is, $\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N(\delta,\ \tau^2)$. The statistical power is then calculated based on this distribution, which is a *conditional* distribution on the unknown treatment difference δ.

The assurance is then defined based on the expected power with a prior distribution on this unknown parameter δ. Using a commonly used conjugated prior normal distribution from previous clinical trials as $\delta \sim N(m,\ v)$, the unconditional distribution can be obtained as $\bar{x}_1 - \bar{x}_2 \sim N(m,\ \tau^2 + v)$. O'Hagan et al (2005) used this formulation and derived the analytical formula for one-sided superiority trial, two-sided superiority trial, non-inferiority trial, and equivalence trial. For example, in a two-sided superiority trial to test the null hypothesis $\mathrm{H_0}$: $\delta = 0$ against the two-sided alternative $\mathrm{H_a}$: $\delta \neq 0$, the null hypothesis is rejected if $|\bar{x}_2 - \bar{x}_1| > \tau Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ significance point of the standard normal distribution. The assurance that the null hypothesis is rejected can be formulated based on the unconditional distribution of $\bar{x}_1 - \bar{x}_2 \sim N(m,\ \tau^2 + v)$ as

$$\gamma = P\left(\text{null hypothesis is rejected}\right) = \mathrm{P}\left(\bar{x}_2 - \bar{x}_1 > \tau Z_{\alpha/2}\right) = \Phi\left(\frac{-\tau\ Z_{\alpha/2} + m}{\sqrt{\tau^2 + v}}\right). \qquad (3)$$

This analytical formulation can reconcile the numerical integration proposed in Chuang-Stein (2006) where the assurance is defined as the probability to produce a successful trial. This article elegantly conceptualizes the assurance from the biopharmaceutical aspects and illustrates the calculations from the designing aspect of a clinical trial. It defines the "success" as a "Trial produces a significant p-value." With this definition, the assurance is given by the Eq. (2) in Chuang-Stein (2006) as follows:

$$\int_{-\infty}^{+\infty} P\left(\text{Trial produces a significant p} - \text{value}|\Delta\right) P\left(\Delta|d\right) d\Delta, \qquad (4)$$

where $P(\text{Trial produces a significant p} - \text{value}|\Delta)$ in Eq. (4) above is in fact the conventional statistical power, which is then "averaged" over the prior distribution of $P(\Delta|d)$ for all possible values of Δ to obtain the assurance. Note that in the formulation defined in Chuang-Stein (2006), as seen in Eq. (4), the notations of Δ and d are equivalent to δ and m in O'Hagan et al (2005). The assurance given in Eq. (4) can be obtained only through a numerical integration and a trapezoid role was used in Chuang-Stein (2006), which was coded in both R and SAS in the appendix.

It can be shown that Chuang-Stein's definition in Eq. (4) is a special case of O'Hagan's definition in Eq. (3) when $\sigma_1^2 = \sigma_2^2 = \sigma^2, n_1 = n_2 = n$ and then $\tau = \sqrt{\frac{2}{n}}\sigma$. The prior variance in O'Hagan et al. (2005), v, is specified as $v = \sqrt{\frac{2}{m}}\sigma$ in Chuang-Stein (2006) (notice that m is the prior sample size in Chuang-Stein but used as the prior mean in O'Hagan et al.). We have programmed this comparison in R (see Appendix A) using O'Hagan et al. (2005) formulation in Eq. (3) (in Appendix A.2) and Chuang-Stein (2006) formulation in Eq. (4) (in Appendix A.1). We reproduced Table 1 in Chuang-Stein (2006) to illustrate the conventional statistical power and assurance calculation when there are 128 and 172 patients per group with prior distribution of $N(2.5,\ (2/m)7.14^2)$ with the prior sample size $m = 25$ and 70.

It can be seen from Appendix A that O'Hagan et al. (2005)'s formulation can be easily implemented using the standard normal cumulative distribution while the Chuang-Stein's formulation will need to call the numerical integration routine in R (i.e., "*integrate*") to obtain the integration in Eq. (4). We reproduced the results in Table 1, which illustrate the difference between the conventional statistical power and assurance. One can use the R code in Appendix A and find that the results from Chunag-Stein's trapezoid numerical integration, the R numerical integration (i.e., "integrate" in Appendix A.1) and O'Hagan et al. (2005) standard normal cumulative distribution (i.e., implemented in R function "pnorm" in Appendix A.2) are exactly the same as seen in Table 1.

Table 1. Assurances calculated with "Known Variance" (in the middle column with boldface fonts), which reproduced Table 1 in Chuang-Stein (2006) by the R code from Appendix A. and the corresponding assurance with "Unknown Variance" from Monte Carlo simulation-based approach (in the right column).

	Sample sizes in prior clinical trials			
	Known variance		Unknown variance	
Sample size in future trials and associated statistical power	$m = 25$	$m = 70$	$m = 25$	$m = 70$
128/group (80% power)	**0.633**	**0.692**	0.627	0.688
172/group (90% power)	**0.677**	**0.756**	0.670	0.752

Notes: To further illustrate the conventional statistical power and assurance under different sample sizes, we provide Fig. 1 and its range of sample sizes for clinical trials. It can be seen and expected that the assurance is typically smaller than the conventional statistical power when sample sizes are larger (>60 in this figure where statistical power >0.5) because the assurance is integrated over all possible parameter vector values. However, it is interesting to observe that when the sample sizes are relatively small and the clinical trials would be underpowered ($<=50\%$), the assurance and the conventional statistical power are similar.

3. Assurance calculations

The illustration in Section 2.3 can only be done using some simple cases with one-dimensional parameter vector θ. Conceptually, assurance defined in Eq. (2) is the expected power to the parameter vector space of θ, which could be high-dimensional. When the expected power involves high-dimensional integration, it will be impractical to obtain the analytical formula to be implemented in statistical software. With the computing technology, we can resolve the assurance computations by Monte Carlo simulation-based approach. Simulation-based computations for designing and analyzing clinical trials have been seen in Kimko and Duffull (2002), Kimko and Peck (2010), and Chow and Chang (2012). Here our focus is for assurance computations. We describe the Monte Carlo simulation-based computations in this section by using R (in Appendix B).

3.1. Bayesian clinical trial simulation (BCTS) to Monte Carlo simulation-based (MCSB) approach

As proposed in O'Hagan et al. (2005) for assurance calculation, the general principle for BCTS is to incorporate sampling from the prior distribution of θ before sampling from the data. Specifically, the general algorithm to compute the assurances of outcomes $A_1, A_2, \ldots, A_k$ is as follows:

1. Define counters I for iteration and $T_1, T_2, \ldots, T_k$ for the assurances, and set all counters to 0. Set the required number, N. Set I = 0 and start looping,
2. Sample θ from the prior distribution,
3. Sample the data and calculate the sufficient statistics using the model and the sampled value of θ from step 2,
4. For $j = 1, 2, \ldots, k$, increment T_j by 1 if the outcome A_j has occurred,
5. Increment I: If $I < N$; go to step 2,
6. For $j = 1, 2, \ldots, k$, estimate assurance $\gamma_j = P(\mathrm{A_j})$ by T_j/N.

In fact, this BCTS can be simplified with the following MCSB approach (hereafter referred as MCSB-General) for computing assurance that involves the following steps:

1. Define counter I for iteration and the required number of simulations, N, (say $N =$ 1,000,000 simulations). Set $I = 0$ and start looping,
2. Sample θ from the joint prior distributions,

3. Calculate the conventional statistical power conditional on this sampled value of θ from step 2 with the data or calculated test statistics using the associated model for hypothesis testing,
4. The assurance can be estimated as the average of the statistical powers from step 3.

We illustrate this BCTS-General to normally distributed data and binary data in the following sections.

3.2. Assurance calculation for normally distributed data when variance is unknown

When the variances are unknown, the commonly used test statistic is the Student t. Under the homogenous variance assumption, this test statistic is formulated as $t = \frac{\bar{x}_2-\bar{x}_1}{\hat{\sigma}\sqrt{\frac{1}{n1}+\frac{1}{n_2}}}$, which follows the Student t-distribution with degrees of freedom, $df = n_1 + n_2 - 2$, where $\hat{\sigma}$ is the estimated pooled standard deviation. In the heterogenous variance assumption, the approximate Satterthwaite $t = (\bar{x}_2 - \bar{x}_1)/\sqrt{\frac{\hat{\sigma}_1^2}{n1} + \frac{\hat{\sigma}_2^2}{n_2}}$ (where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the estimated sample variances) is used with degrees of freedom matching the moments (see Chen and Peace, 2011, for details).

The standard two-sided test for the null hypothesis of no treatment difference H_0: $\delta = 0$ against the two-sided alternative H_a: $\delta \neq 0$ is to reject the null hypothesis if $|t| > t_{\alpha/2,df}$. The statistical power can then be calculated based on this t-distribution. This distribution can also be used to calculate the assurance by two-dimensional numerical integration over the parameter space of δ and σ^2 with the non-central t-distribution. However, the Monte Carlo simulation-based BCTS approach can be easily implemented for this purpose. Corresponding to the general algorithm in Section 3.1, this approach (hereafter referred as MCSB-Normal) can be implemented in the following steps:

1. Set counter $I = 0$ and the number of simulations, N (say, $N = 1{,}000{,}000$),
2. Sample δ and σ^2 from their joint prior distribution,
3. Sample $\bar{x}_2 - \bar{x}_1 \sim N(\delta,\ (n_1^{-1} + n_2^{-1})\sigma^2)$ and $(n_1 + n_2 - 2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{df}$, calculate the t-test statistic and statistical power,
4. Estimate the assurance with the average of the resulted sample of N statistical powers.

This MCSB-Normal approach is implemented in R as seen in Appendix B. We first implemented this MCSB-Normal in Appendix B.1 with known variance (i.e., the function "ANDks" in short for "Assurance for Normal Data with Known Sigma") to confirm the results given by Chuang-Stein (2006) shown in the middle column of Table 1. We then programed the MCSB-Normal with unknown variance in Appendix B.2 (i.e., ANDus in short for "Assurance for Normal Data with unknown sigma") for the same scenarios and the estimated assurances from this MCSB-Normal shown in the right column in Table 1. It can be seen from these results that the assurances with unknown variance are smaller than the assurances with known variance. This is considered reasonable and consistent with general conclusion in the statistical power. When the variance is known, then the calculations can be done more precisely, whereas when the variance is unknown, it needs to be estimated and therefore introduces additional variability in assurance calculations.

3.3. Assurance calculation for binary data with O'Hagan et al.'s (2005) formulation

In clinical trials with binary data with x_i successes from total n_i patients for treatment i ($i = 1, 2$), denote p_i as the population success rate for treatment i. Then the null hypothesis to test the treatment efficacy is H_0: $p_1 = p_2$. The classical statistical test is based on approximated normality of the sample proportions $\hat{p}_i = x_i/n_i$ (see, e.g., in Chen and Peace, 2011) and the null hypothesis is rejected in a two-sided test if $|Z| > Z_{\alpha/2}$ where $Z = (\hat{p}_2 - \hat{p}_1)/\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ is approximately standard normally distributed. The conventional statistical power can then be approximated by

$$P\left(R|p_1, p_2\right) \approx \Phi\left(-Z_{\alpha/2} + \frac{p_2 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right), \tag{5}$$

where $\phi(.)$ is the standard normal density function.

This definition of statistical power is a *conditional* probability conditional on two unknown parameters p_1 and p_2 from both treatments. The assurance would then be calculated by integrating these two unknown parameters p_1 and p_2 from their joint prior distributions, which are typically beta conjugate prior distributions. This would not be feasible to obtain analytical formula based on Eq. (5). However, this can be easily implemented with MCSB approach (hereafter referred as MCSB-Binary) as follows:

1. Set counter $I = 0$ and the number of simulations N (say, $N = 1{,}000{,}000$),
2. Sample p_1 and p_2 from their prior distributions that are typically beta-distributions,
3. Calculate the Z-statistic and then the statistical power using Eq. (5),
4. Estimate the assurance with the average of the resulted sample of N statistical powers.

This MCSB-Binary approach is implemented in R as seen in Appendix B.3. We make use of Example 4 in sec. 5.1 from O'Hagan et al. (2005) for clinical trials in rheumatoid arthritis where prior information for control drug is specified as $\mathrm{E}(p_1) = 0.2$ and $\mathrm{sd}(p_1) = 0.08$ from the published results for methotrexate in Kremer et al. (2002). This prior information can be represented by a beta-distribution as $p_1 \sim \mathrm{Beta}(5, 20)$. The new drug has more uncertainty as specified by $\mathrm{E}(p_2) = 0.4$ and $\mathrm{sd}(p_2) = 0.17$, which corresponds to beta-distribution as $p_2 \sim \mathrm{Beta}(3, 4.5)$.

Recognizing the ineffectiveness, the development team used a weighting scheme with a 0.15 probability for Beta (2, 23) and 0.85 for Beta (3, 4.5). The trial was planned with unequal sample size, $n_1 = 200$ patients in the control (methotrexate) group and $n_2 = 400$ in the treatment group for a two-sided 5% significance level test for superiority. With these sample sizes to detect an improvement from $p_1 = 0.2$ to $p_2 = 0.3$ (i.e., a 50% treatment effect), the statistical power can be calculated by Eq. (5) to be 78%. Using the MCSB-Binary described above, which is implemented in Appendix B.3, the assurance is 0.633, smaller than the conventional power of 0.78.

3.4. Assurance calculation for binary data with logit-normal formulation

In designing and analyzing binary data, it is common to consider the logit-normal transformation to the response rates that then lead to the logistic regression for categorical data. The logit-normal formulation was introduced in Mead (1965) and Aitchison and Shen (1980). In analyzing dose–response relationship incorporating historical control data, Chen (2010) used this logit-normal formulation with an empirical Bayes approach.

Traditionally, this transformation is as follows: $logit(p_i) = log(\frac{p_i}{1-p_i}) = \alpha_i$ for each treatment $i = 1$ and 2. It is known that the ratio $p_i/(1 - p_i)$ is the odds of success so that $\mathrm{logit}(p_i)$ is often called the log odds, which is used as logit link in generalized linear model for binary or binomial data (see, e.g., Chen and Peace, 2011). Transforming this back, we would have $p_i = \frac{e^{\alpha_i}}{1+e^{\alpha_i}}$; therefore, the typical null hypothesis H_0: $p_1 = p_2$ is equivalent to H_0: $\alpha_1 = \alpha_2$.

With this logit-normal transformation to p_{i}, the α_{i} is usually assumed to be normally distributed as

$$\alpha_i \sim N\left(log\left(\frac{p_i}{1-p_i}\right), \sigma_i^2\right) \tag{6}$$

and then the MCSB approach can be implemented to sample the α_i from the above distribution in Eq. (6).

The MCBS approach (hereafter referred as MCSB-Power) to estimate the statistical power for testing treatment efficacy between two treatments with sample sizes n_1, n_2 and probabilities p_1 and p_2, can be implemented by the following steps:

1. Set counter $I = 0$ and the number of simulations, N (say, $N = 1{,}000{,}000$),
2. Sample $x_i \sim Binomial(n_i, p_i)$ and perform logistic regression to test the null hypothesis H_0: $\alpha_1 = \alpha_2$, or essentially H_0: $p_1 = p_2$,
3. If the associated p-value from this logistic regression is less than 0.05 for H_0, increase I by one,
4. Repeat steps 2 and 3 for N times and the statistical power can be then estimated by I/N.

Now this MCSB approach can be similarly implemented to calculate the assurance using the logit-normal formulation (hereafter referred as "MCSB-LogitNormal") for testing treatment efficacy between two treatments with sample sizes n_{i}, probabilities p_i and σ_i^2 ($i = 1{,}2$), can be implemented by the following steps:

1. Set counter $I = 0$ and the number of simulations, N (say, $N = 1{,}000{,}000$),
2. Sample $\alpha_i \sim N(log(\frac{p_i}{1-p_i}), \sigma_i^2)$ from Eq. (6) and then calculate $p_i = \frac{e^{\alpha_i}}{1+e^{\alpha_i}}$ using the sampled α_i,
3. Sample $x_i \sim Binomial(n_i, p_i)$ and perform logistic regression to test the null hypothesis H_0: $\alpha_1 = \alpha_2$, or essentially H_0: $p_1 = p_{2,}$
4. If the associated p-value from this logistic regression is less than 0.05 for H_0, increase I by one,
5. Repeat steps 2 and 4 for N times and the statistical power can be then estimated by I/N.

Both MCSB-Power and MCSB-LogitNormal proposed above can be easily implemented in any software that handles logistic regression. Appendix C is the R code for these two approaches.

To illustrate these approaches, again we make use of sec. 5.1 in O'Hagan et al. (2005) for clinical trials in rheumatoid arthritis where $p_1 = 0.2$ and $p_2 = 0.3$, which corresponds to $\alpha_1 = -1.386$ and $\alpha_2 = -0.847$, respectively. To simplify the illustration and without loss of generality, we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and consider one case of small variance 0.01 and the other case of large variance 0.1. In addition, we consider equal sample size for the two treatments (i.e., $n = n_1 = n_{2)}$. With these settings, we run these two MCSB approaches for sample sizes from 50 to 800 and calculate the statistical power using MCSB-Power and assurance using MCSB-LogitNormal. The results are summarized in Table 2 and graphically displayed in Fig. 2.

It can be seen from Table 2 and Fig. 2 that the statistical power is generally larger than the assurance for sufficient large sample size (>180 in this simulation) but similar for small sample size (< 180 in this simulation). This conclusion is consistent with the results in Table 1 and Fig. 1. It is observed that the statistical power is greater than 0.8 for sample size 300 per group. This power is similar to the result in Section 3.3 where it is 0.78 for sample size of 200 for the control group and 400 for the treatment group. It is also observed that given this sample of 300, the assurance (expected probability to have a successful trial) is 0.76 and 0.67 for variance = 0.01 and variance = 0.1, respectively, which are smaller but not by much.

Table 2. "MCSB-Power" approach for statistical power (the 2nd column) and "MCSB-LogitNormal" approach for assurance with variance = 0.01 (the 3rd column) and assurance with variance = 0.1 (the 4th column) for various sample sizes (1st column).

Sample size /group ($n = n_1 = n_2$)	Statistical power	Assurance with variance = 0.01	Assurance with variance = 0.1
50	0.195	0.212	0.262
100	0.382	0.377	0.434
200	0.640	0.616	0.580
300	0.809	0.761	0.665
400	0.907	0.839	0.702
500	0.957	0.899	0.737
600	0.980	0.925	0.767
700	0.991	0.945	0.786
800	0.996	0.955	0.801

In order to have assurance to be 0.8 to ensure an absolute successful trial, the sample size would have to be 350 (where the power is 0.85) for smaller prior variance (i.e., when variance = 0.01) and an enormous 800 (where the power is 0.99) for large prior variance (i.e., when variance = 0.1), as seen in Table 2 and Fig. 2.

With a further increase in the sample size from 300 to 400, the associated statistical power increases from 0.8 to 0.9, roughly a 12% increase. Nonetheless, the assurance increase is about 10% from 0.76 to 0.84 for smaller variance of 0.01 and 6% from 0.66 to 0.70 for large variance of 0.1. It is important to observe that the assurance is bounded. In our case with the larger variance of 0.1, the assurance limit is roughly 0.7, no matter how big the sample size is. This indicates that assurance is constrained by the variability of prior information, which is reasonable because with uncertainty in the prior distribution and without collecting the actual data, it is impossible to be 100% assured that a trial will be successful. In light of this, assurance should be compared within the context, rather than in a vacuum or using the absolute magnitude.

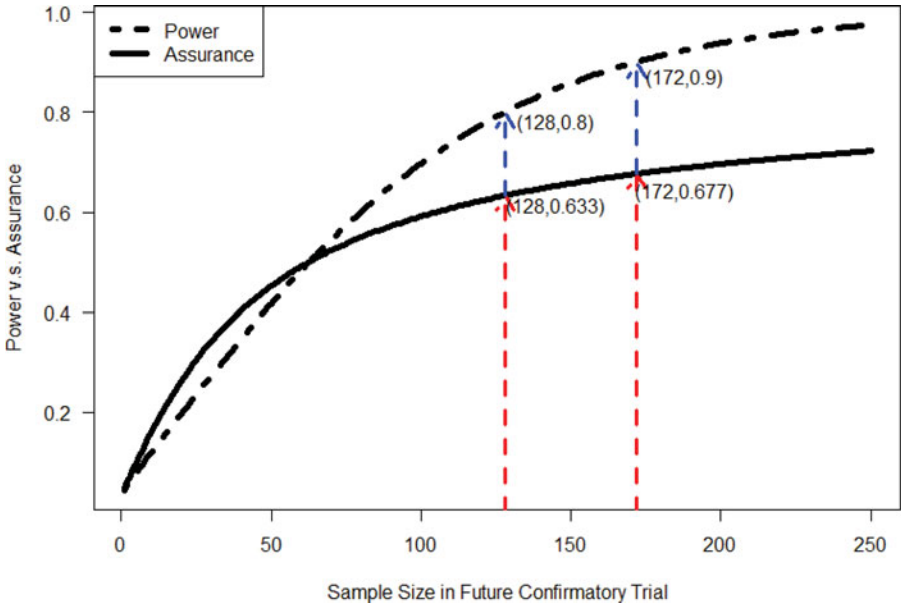

Figure 1. Statistical power and assurance from different sample sizes for clinical trials. The two vertical arrow lines correspond to the sample sizes of 128 and 172 in Table 1 where the statistical powers are 0.8 and 0.9, and assurances are 0.633 and 0.677, respectively.

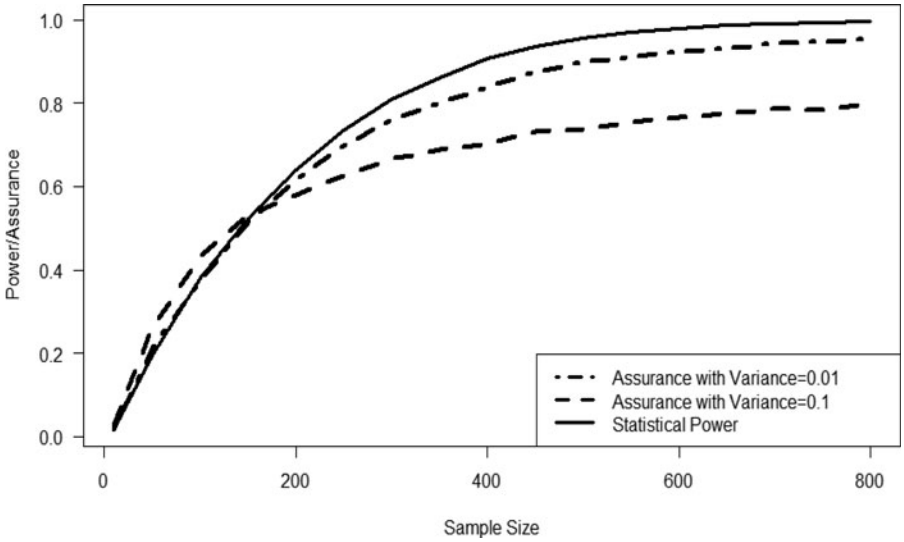

Figure 2. Statistical power and assurance for a range of sample sizes under two different variances of 0.01 and 0.1.

4. Discussion

In this article, we illustrated the transition from the conventional statistical power to assurance proposed in O'Hagan and Stevens (2001) in designing clinical trials. The conventional statistical power is the probability of rejecting the null hypothesis conditional on the specified treatment effect, whereas the assurance is the unconditional probability of a successful clinical trial averaged over the parameter space of this pre-specified treatment effect. Then the calculation of the assurance involved a high-dimensional integration would have to resort to numerical integration. We promote the Monte Carlo simulation-based approach in this article and illustrated its implementation in R for clinical trials with normally distributed data given known or unknown variances, as well as clinical trials with binary data from beta and logit-normal distributions.

It is common knowledge that a traditionally powered clinical trial at 80% to 90% does not guarantee 80% to 90% of probability of success, as the power calculation is based upon a pre-specified fixed treatment effect, which most likely will be different from the true treatment effect. Typically, the assurance is lower than the statistical power for a sufficient sample size, even though we observe that the assurance could be higher than the statistical power for underpowered clinical trials as seen in Figs. 1 and 2. It is well known that as the sample size increases and approaches to infinity, the traditional statistical power will approach to 1. However, the assurance will be bounded by a value less than 1 even when the sample size approaches to infinity. This can be analytically illustrated using the descriptions in Section 2.3. As study sample sizes (e.g., n_1 and n_2 in our examples) approach to infinity, the standard error (i.e., τ) should approach to zero. The assurance defined in O'Hagan et al. (2005) would then approach to $\gamma = \Phi(\frac{m}{\sqrt{v}})$, which is the prior probability of success (positive outcomes).

In the real world of clinical trials, we believe that the assurance can provide a more realistic and robust measure of probability of success than the traditional power can. Assurance does depend on the prior distribution of treatment effects, which can be very subjective with varying Bayesian priors. Thus, the related issues in the Bayesian approach also apply to assurance. Nevertheless, assurance typically will be used by the pharmaceutical industry sponsors for internal decision-making and therefore it is the industry sponsor’s risk of using it. As such, it will also be at the industry sponsor’s best interest to get the most relevant prior distribution for the treatment effect. This can be accomplished by various ways such as historical data driven or expert elicitation or a hybrid of both. Good decision-making is important for clinical development and therefore any method that can enhance good decision-making will be beneficial to patients, sponsors and regulators, as well as society in general.

Acknowledgments

The authors would like to thank the anonymous referee and Drs. Naitee Ting and Joseph C. Cappelleri for their constructive comments and suggestions.

Appendix A: R code to compare Chuang-Stein’s (2006) numerical integration formulation with O’Hagan et al.’s (2005) formulation when prior variance is known

```
########################################################################
# A.1. Use R “integrate ” instead of the trapezoidal numerical integration # in Chuang-Stein
(2006), which can compute the assurance faster
########################################################################
sprob.Chuang = function(prior.mean,prior.sd,prior.size,post.size){
prior.sdm = sqrt(2/prior.size)*prior.sd # prior sd for the mean
post.sdm = sqrt(2/post.size)*prior.sd # posterior sd for the mean
# fn for the Prob(trial produces a significant p-val)*prior distribution
integrand ← function(delta)
pnorm(1.96*post.sdm,mean = delta,sd = post.sdm,lower.tail = FALSE,log.p = FALSE)*
    dnorm(delta,mean = prior.mean,sd = prior.sdm,log = FALSE)
# Numerical integration of delta from -Inf to Inf
avg = integrate(integrand, lower = -Inf, upper = Inf)$value
# output
avg
} # end of “sprob.Chuang”
## Run the function and Reproduce Table I in Chuang-Stein (2006)
> sprob.Chuang(2.5,7.14,25,128)
[1] 0.6330728
> sprob.Chuang(2.5,7.14,25,172)
[1] 0.6767027
> sprob.Chuang(2.5,7.14,70,128)
[1] 0.6915049
> sprob.Chuang(2.5,7.14,70,172)
[1] 0.7555993
######################################################################
# A.2: O’Hagan’s formulation to call “pnorm”
######################################################################
sprob.OHagan = function(prior.mean,prior.sd,prior.size,post.size){
tau = sqrt(2/post.size)*prior.sd; v = sqrt(2/prior.size)*prior.sd
  sprob.OHagan = pnorm((qnorm(0.025)*tau+prior.mean)/sqrt(tau^2+v^2))
# output
sprob.OHagan
} # end of “sprob.OHagan”
## Run the code and Reproduce Table I in Chuang-Stein with O’Hagan et al.
> sprob.OHagan(2.5,7.14,25,128)
[1] 0.6330783
> sprob.OHagan(2.5,7.14,25,172)
[1] 0.6767073
> sprob.OHagan(2.5,7.14,70,128)
[1] 0.6915124
> sprob.OHagan(2.5,7.14,70,172)
[1] 0.7556054
```

Appendix B: R code for BCTS

```
#############################################################################
# B.1: Assurance for Normal Data with Known Sigma (ANDks)
# to check with the results from Appendix A
#############################################################################
ANDks = function(nsimu,prior.mean,prior.sd,prior.size,post.size){
sim.pow = rep(0, nsimu)
for(i in 1:nsimu){
# calculate the standard deviation for the means
prior.sdm = sqrt(2/prior.size)*prior.sd # prior sd for the mean
post.sdm = sqrt(2/post.size)*prior.sd # posterior sd for the mean
# sample the prior
Delta = rnorm(1,prior.mean,prior.sdm)
# with the sampled prior, calculate the power
sim.pow[i] = pnorm(qnorm(1-alpha/2)*post.sdm,
mean = Delta,sd = post.sdm,lower.tail = FALSE,log.p = FALSE)
} # end of i-loop
# average the simulated power
mean(sim.pow)
} # end of “ANDks” function
## run the code to check with the calculations in Appendix A
> ANDks(1000000,2.5,7.14,25,128)
[1] 0.6329084
> ANDks(1000000,2.5,7.14,25,172)
[1] 0.6761969
> ANDks(1000000,2.5,7.14,70,128)
[1] 0.6915248
> ANDks(1000000,2.5,7.14,70,172)
[1] 0.7555312
##############################################################################
# B.2: Assurance for Normal Data with unknown sigma(ANDus)
##############################################################################
ANDus = function(nsimu,prior.mean,prior.sd,prior.size,post.size){
sim.pow = rep(0, nsimu)
for(i in 1:nsimu){
```

```
# sample chisq for sigma since (n-1)*s^2/sigma^2 ~chisq(n-1)
sd = sqrt((prior.size-1)*prior.sd^2/rchisq(1,df = prior.size-1))
# calculate the standard deviation for the mean
prior.sdm = sqrt(2/prior.size)*sd # prior sd for the mean
post.sdm = sqrt(2/post.size)*sd # posterior sd for the mean
# sample the prior
Delta = rnorm(1, prior.mean,prior.sdm)
# with the sampled prior, calculate the power
sim.pow[i] = pnorm(1.96*post.sdm,mean = Delta,
sd = post.sdm,lower.tail = FALSE,log.p = FALSE)
} # end of i-loop
# assurance is the average of simulated power
mean(sim.pow)
} # end of "ANDus" function
#### run the function for assurance with unknown variance
> ANDus(1000000,2.5,7.14,25,128)
[1] 0.627181
> ANDus(1000000,2.5,7.14,25,172)
[1] 0.6700293
> ANDus(1000000,2.5,7.14,70,128)
[1] 0.687614
> ANDus(1000000,2.5,7.14,70,172)
[1] 0.7515066
##################################################
# B.3: Assurance with binary clinical trial in O'Hagan et al. Example 4
##################################################
library(Rlab) # for rbern
nsimu = 1000000;alpha = 0.05; post.size1 = 200; post.size2 = 400
sim.pow = rep(0,nsimu)
for (i in 1:nsimu){
# sample Beta for p1
p1 = rbeta(1,5,20)
# sample Beta for p2 from a mixture
w = rbern(1,0.15);p2 = w*rbeta(1,2,23)+(1-w)*rbeta(1,3,4.5)
# z-value in equation
z.val = (p2-p1)/sqrt(p1*(1-p1)/post.size1+p2*(1-p2)/post.size2)
# with the sampled prior, calculate the power
sim.pow[i] = pnorm(-qnorm(1-alpha/2)+z.val)
} # end of i-loop
# Assurance as the mean
mean(sim.pow)
[1] 0.6334372
```

Appendix C: MCSB approach for statistical power and assurance

```
#######################################
# MCSB function
#######################################
```

```
pow2assurance = function(nsimu,n1,n2,pA,pB,sig2A,sig2B,sig2A2,sig2B2,alpha){
# initializes the power and assurance
pow = assu1 = assu2 = 0
# loop for calculation
for(i in 1:nsimu){
### power simulation
xA = rbinom(n1,1,pA);xB = rbinom(n2,1,pB)
dd = data.frame(x = c(xA,xB),trt = c(rep(“A”,n1),rep(“B”,n2)))
md = glm(x~trt,dd,family = “binomial”);
pval.md = summary(md)$coef[“trtB”,“Pr(>|z|)”]
pow = pow+sum(pval.md < alpha)
### assurance simulation for sigma1
alphaA = rnorm(1,log(pA/(1-pA)),sqrt(sig2A));
alphaB = rnorm(1,log(pB/(1-pB)),sqrt(sig2B))
pAs = exp(alphaA)/(1+exp(alphaA));pBs = exp(alphaB)/(1+exp(alphaB));
xA = rbinom(n1,1,pAs);xB = rbinom(n2,1,pBs)
dd = data.frame(x = c(xA, xB), trt = c(rep(“A”, n1), rep(“B”, n2)))
md = glm(x~trt, dd, family = “binomial”);
pval.md = summary(md)$coef[“trtB”,“Pr(>|z|)”]
assu1 = assu1+sum(pval.md < alpha)
### assurance simulation for sigma2
alphaA = rnorm(1,log(pA/(1-pA)),sqrt(sig2A2));
alphaB = rnorm(1,log(pB/(1-pB)),sqrt(sig2B2))
pAs = exp(alphaA)/(1+exp(alphaA));pBs = exp(alphaB)/(1+exp(alphaB));
xA = rbinom(n1,1,pAs);xB = rbinom(n2,1,pBs)
dd = data.frame(x = c(xA, xB), trt = c(rep(“A”, n1), rep(“B”, n2)))
md = glm(x~trt, dd, family = “binomial”);
pval.md = summary(md)$coef[“trtB”,“Pr(>|z|)”]
assu2 = assu2+sum(pval.md < alpha)
} #End of i-loop
# output
list(pow = pow/nsimu,assu1 = assu1/nsimu,assu2 = assu2/nsimu )
}# end of pow2assurance
```

ORCID

Ding-Geng (Din) Chen http://orcid.org/0000-0002-3199-8665
Shuyen Ho http://orcid.org/0000-0001-8356-1089

References

Aitchison, J., Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* 67(2):261–272.

Barry, S. M., Carlin, B. P., Lee, J. J., Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Chen, D. G. (2010). Incorporating historic control information with Empirical Bayes. *Journal of Computational Statistics and Data Analysis* 54:1646–1656.

Chen, D. G., Peace, K. E. (2011). *Clinical Trial Data Analysis Using R*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Chow, S. C., Chang, M. (2012). *Adaptive Design Methods in Clinical Trials*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Chow, S. C., Wang, H., Shao, J. (2003). *Sample Size Calculations in Clinical Research*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics* 5:305–309.

Kimko, H., Duffull, S. B. (2002). *Simulation for Designing Clinical Trials: A Pharmacokinetic-Pharmacodynamic Modeling Perspective*. Boca Raton, FL: Chapman and Hall/CRC.

Kimko, H. C., Peck, C. C. (2010). *Clinical Trial Simulations: Applications and Trends (AAPS Advances in the Pharmaceutical Sciences Series)*. New York: Springer.

Kremer, J. M., Genovese, M. C., Cannon, G. W., Caldwell, J. R., Cush, J. J., Furst, D. E., Luggen, M. E., Keystone, E., Weisman, M. H., Bensen, W. M., Kaine, J. L., Ruderman, E. M., Coleman, P., Curtis, D. L., Kopp, E. J., Kantor, S. M., Waltuck, J., Lindsley, H. B., Markenson, J. A., Strand, V., Crawford, B., Fernando, I., Simpson, K., Bathon, J. M. (2002). Concomitant leflunomide in patients with active rheumatoid arthritis despite stable doses of methotrexate: A randomized double blind placebo controlled trial. *Annals of Internal Medicine* 137:726–733.

Mead, R. (1965). A generalised logit-normal distribution. *Biometrics* 21(3):721–732.

O'Hagan, A., Stevens, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making 2001* 21:219–230.

O'Hagan, A., Stevens, J. W., Campbell, M. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics* 4:187–201.

Ren, S., Oakley, J. E. (2014). Assurance calculations for planning clinical trials with time-to-event outcomes. *Statistics in Medicine* 33:31–45.

Spiegelhalter, D. J., Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. *Statistics in Medicine* 5:1–13.

Spiegelhalter, D. J., Abrams, K. R., Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: Wiley.

Walter, Y., Chen, D. G. (2014). *Clinical Trial Biostatistics and Biopharmaceutical Applications*. Boca Raton, FL: Chapman and Hall/CRC.