Assurance in Intervention Research: A Bayesian Perspective on Statistical Power

Ding-Geng Chen *University of North Carolina at Chapel Hill*
Mark W. Fraser *University of North Carolina at Chapel Hill*
Gary Cuddeback *University of North Carolina at Chapel Hill*

ABSTRACT ***Objective:*** This article introduces Bayesian assurance as an alternative to traditional power analysis in intervention research. Bayesian assurance is defined as the unconditional probability of identifying an intervention effect. ***Method:*** Assurance can be calculated as the expected statistical power based on a prior distribution of the unknown parameters related to the effect size. Using Monte Carlo simulation methods, we demonstrate Bayesian assurance in 2 small-scale randomized trials: a trial of motivational interviewing for patients with behavioral health disorders and a trial of a specialty mental health probation. ***Results:*** The findings suggest that traditional statistical power is highly sensitive to misspecification. Because assurance can be calculated across all possible effect sizes, it controls the uncertainty associated with the selection of a point effect size in traditional power estimation. Assurance usually produces larger sample-size estimates, and thus cutoff values for assurance may be lower than those typically used in classical power estimation. ***Conclusions:*** Compared to traditional power estimation, assurance appears to be more robust against inaccurate prior information. Assurance may be a preferred method for estimating sample sizes when prior information is poor and the costs of underpowering a study are great.

KEYWORDS: assurance, power, intervention research, Bayesian, sample size, effect size

doi: 10.1086/696239

A defining feature of intervention research in the health and social services is a design and development process in which one study informs the next study. When successful, small pilot or feasibility studies often lead to efficacy and effectiveness trials that, in turn, lead to cost-effectiveness and cost-benefit studies (Fraser & Galinsky, 2010; Fraser, Richman, Galinsky, & Day, 2009). In this process, determining the appropriate number of participants (i.e., sample size) for each sequential trial is a practical challenge because, in the early stages of intervention development, researchers often do not have good information on which to

base effect size estimates. Because of the cost of trials and the implications of findings for future studies, making efficient and accurate power calculations has high practical value.

Statistical power is traditionally defined as the probability of rejecting the null (i.e., identifying an effective intervention) if the true intervention effect equals a specific value or effect size. Therefore, statistical power is a conditional probability related to an expected value defined in the research design process (Cohen, 1988). In practice, this value is typically obtained from pilot studies and previous research, including systematic reviews and meta-analyses. However, the effect size value can also be based on knowledge and prior experience, which could be very different from the true unknown intervention effect size. In addition to leading to unnecessarily large samples, errors in estimating an expected effect size can lead to sample sizes that are too small, and this can result in studies with inadequate power. An inadequately powered study can produce findings that influence the decision to cease development of an intervention, whereas a study with adequate statistical power might have shown the intervention as promising.

Because of this imprecision, traditional statistical power analysis has been criticized for failing to adequately consider the uncertainty associated with specifying an expected effect size (Du & Wang, 2016). To calculate power in such a way that it is less reliant on a single point effect size estimate, Bayesians developed the concept of *assurance*. Assurance is defined as the unconditional probability of rejecting the null hypothesis (Chen & Ho, 2016; Chuang-Stein, 2006; O'Hagan & Stevens, 2001; O'Hagan, Stevens, & Campbell, 2005; Ren & Oakley, 2014). Assurance can be obtained as the expected power with respect to the distribution of a pre-observed effect size along with related parameters. This framework leads to the Bayesian paradigm. Conceptualized this way, assurance is sometimes called *Bayesian assurance*.

Bayesian assurance, as an alternative to traditional statistical power, is a developing concept in social and health sciences research. The purpose of this paper is to introduce assurance and demonstrate its computation with Monte Carlo simulations.

Conventional Statistical Power and Bayesian Assurance

In addition to the design of new interventions, the general objective of intervention research is to test whether a new intervention is superior to treatment as usual. To demonstrate the effectiveness of a new intervention, researchers determine how many participants should be enrolled in each intervention arm of a trial (i.e., the intervention or treatment arm and the treatment-as-usual or control arm). Typically, a sample size estimate is based on a power calculation. Translated into statistical terms, the null hypothesis (H_0) for intervention research is defined as the two arms being no different versus the alternative hypothesis (H_a), defined as the new intervention being superior to treatment as usual (i.e., a one-tailed test). Statistical power (π) is then defined as the probability of rejecting the null hypothesis when it

is false. Power is typically set at 0.80. The associated sample size can then be determined based on this power and the Type I error rate.

Following the notation from O'Hagan et al. (2005) and from Chen and Ho (2016), we denote R as the event of rejecting the null hypothesis (i.e., outcomes appear to be in favor of a new intervention). Conventional statistical power can be then written as follows:

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

where $\pi(\theta)$ is the power function and θ is a parameter vector including the treatment effect, sample variance, and all possible other parameters. As defined in Equation 1, power is the probability of R conditioned on a set of unknown parameters in vector θ. The value of this *conditional* probability, as well as the associated sample-size calculation, is then dependent on the unknown parameter vector θ.

Generally, parameter θ cannot be known precisely in practice, and it is rare that observed data from systematic reviews, meta-analyses, and pilot studies will provide fully accurate estimates (Chen & Peace, 2013). Therefore, power estimation—as one of the more important activities in intervention research—can lead to either overpowering or underpowering an intervention study. Typically, overpowering means that resources are likely wasted by recruiting more participants than needed. However, underpowering is potentially worse. A study that is underpowered can produce nonsignificant findings that mislead stakeholders into thinking that an intervention, if taken to scale, will have no added benefit over treatment as usual.

Reformulating the Concept of Power: Assurance

Bayesians are reformulating the core idea of power in research design to address the risk of overpowering or underpowering. Spiegelhalter and Freedman (1986) proposed the concept of assurance as a predictive approach to estimating sample sizes in clinical trials. Acknowledging that prior information has always been used to calculate sample sizes, Spiegelhalter, Abrams, and Myles (2004) later conceptualized assurance as a hybrid frequentist–Bayesian concept. O'Hagan and Stevens (2001) also advanced assurance, naming the concept Bayesian assurance (denoted by γ). They conceptualized assurance as an alternative to statistical power—defining it as an *unconditional* probability—by integrating all prior parameters (θ) to reject the null hypothesis; that is, $\gamma = P(R)$, where R is rejection of the null hypothesis to confirm that a new intervention is superior to treatment as usual.

Assurance can then be defined as the expected power in the parameter space of θ. It can be seen that

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

where the expectation is related to the prior distribution, $P(\theta)$.

With this definition, Bayesian assurance extends the frequentist approach to statistical power by averaging (or integrating out) its conditionality across all plausible prior values of θ. Assurance can then provide an unconditional probability of the success of a clinical trial regardless of prior parameters that might not be well observed in the design and development of a new intervention. When derived in this way, assurance is more realistic and robust than estimates based on conventional statistical power calculations.

Assurance Calculations Using Monte Carlo Simulation

Conceptually, Bayesian assurance as defined in Equation 2 is the expected power in the parameter space of θ. Assurance is calculated by averaging/integrating across all possible parameters in the space of θ. Of course, this integration may be high dimensional given that θ could include the intervention effect size along with other related parameters (e.g., covariates). When the expected power involves high-dimensional integration, attempting to solve the analytical formula with routine statistical software is impractical because the computations require high-capacity computing. However, with routine computing technology, analysts can resolve the Bayesian assurance computations using a Monte Carlo simulation-based approach. Monte Carlo simulation is commonly used in designing and analyzing clinical trials in medicine and pharmaceutical studies (Chen & Chen, 2017; Kimko & Peck, 2010).

Detailed implementation in the open-source software R (which can be freely downloaded from http://www.r-project.org) can be found in Chen and Ho (2016). For readers who are interested in reproducing the calculations in this paper and using assurance in their own work, the R program can be requested from the authors. These are the steps to estimate Bayesian assurance using Monte Carlo simulation:

1. Define Counter I for iteration and the required number of simulations, N (e.g., $N = 1{,}000{,}000$). (Note, however, that other than the pragmatic limitations inherent in available computing power, there are few guidelines for the size of N. Typically, N is recommended to be sufficiently large so as to stabilize the assurance distributions.) Set $I = 0$ and start looping.
2. Sample θ from the joint prior distributions.
3. Calculate 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. Estimate Bayesian assurance by averaging statistical power from Step 3.

In the next section, we have illustrated a Monte Carlo simulation-based approach to estimate Bayesian assurance. We focused on continuous and binary data be-

cause they are commonly used in research studies; however, other types of data can be easily implemented with some modifications in Steps 2 and 3.

Assurance Calculation for Normally Distributed Data with Unknown Variances

When variances are unknown, a commonly used statistic is the *t*-test. Under the homogeneous 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. Under the heterogeneous variance assumption, the Satterthwaite approximation—

$$t = (\bar{x}_2 - \bar{x}_1) \Big/ \sqrt{\frac{\hat{\sigma}_1^2}{n1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

(where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the estimated sample variances)—is used with degrees of freedom (*df*) matching the moments (see Chen & Peace, 2011; Chen, Peace, & Zhang, 2017).

The standard test for the null hypothesis of no intervention effect—H_0: $\delta = 0$ against the one-sided alternative H_a: $\delta > 0$—is to reject the null hypothesis if $t > t_{\alpha,df}$. Statistical power can then be calculated based on this *t*-distribution. This distribution can also be used to calculate a Bayesian assurance by two-dimensional numerical integration over the parameter space of δ and σ^2 with the noncentral *t*-distribution. However, the Monte Carlo simulation-based approach is more easily implemented in this setting. Corresponding to the steps previously described, this approach (hereafter referred to as *MC-Normal*) can be implemented in the following steps:

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

Assurance Calculation for Binary Data

In any intervention research that produces binary data with x_i successes from total participants (n_i) for intervention i ($i = 1, 2$), denote p_i as the population success rate for intervention i. Then the null hypothesis to test the intervention effectiveness 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., Chen & Peace, 2011; Chen et al., 2017). The null hypothesis is rejected in a one-sided test if $Z > Z_\alpha$, where

$$Z = (\hat{p}_2 - \hat{p}_1) \Big/ \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is approximately standard normally distributed. The conventional statistical power can then be approximated by

$$\pi(R|p_1, p_2) \approx \Phi\left(-Z_\alpha + \frac{p_2 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right), \tag{3}$$

where Φ is the standard normal density function.

This definition of statistical power is a *conditional* probability that is based on two unknown parameters (p_1 and p_2) from both treatments. By definition in Equation 2, Bayesian assurance would then be calculated by integrating these two unknown parameters (p_1 and p_2) from their joint prior distributions. This integration would not be feasible using the analytical formula in Equation 3. However, the calculation can be implemented easily with a Monte Carlo simulation-based approach (hereafter referred to as *MC-Binary*) in the following steps:

1. Set Counter I = 0 and the number of simulations N (e.g., N = 1,000,000).
2. Sample p_1 and p_2 from their prior distributions.
3. Calculate the Z-statistic and then the associated statistical power using Equation 3.
4. Estimate the assurance with the average of the resulted sample of N statistical powers.

Applications of MC-Normal and MC-Binary

We illustrate both the MC-Normal and MC-Binary approaches in this section. The first example is from Martino, Carroll, Nich, and Rounsaville (2006)—a report of a randomized controlled pilot study of motivational interviewing for individuals with psychotic and drug-use disorders. We used *mean days of alcohol use from the past 28 days* as a continuous outcome. The second example comes from a small-scale randomized controlled trial of a specialty probation program for people with serious mental illness. For this analysis, we used a *technical probation violation* as a binary outcome (i.e., presence or absence of probation violation).

A Randomized Controlled Pilot Study of Motivational Interviewing

As described in Martino et al. (2006), a randomized controlled pilot study was conducted to examine the efficacy of a motivational interviewing intervention implemented over two sessions and adapted for people dually diagnosed with psychotic

and substance-use disorders (i.e., dual diagnosis motivational interviewing; DDMI). DDMI was compared with a standard psychiatric interview (SI) conducted over two sessions. For this study, 44 treatment-seeking participants were recruited and randomly assigned to the two treatment conditions; 24 of the 44 participants were randomized to receive DDMI, and 20 participants were randomized to the standard intake. The primary outcomes were "days of primary drug use, secondary drug use, alcohol use and psychotropic medication adherence, proportion of participants admitted into the program and days of attendance" (Martino et al., 2006, p. 1479). Participants were followed and assessed at four time points: baseline, and 4-, 8- and 12-week follow-up. Martino et al. concluded that both DDMI and SI showed improved treatment outcomes, but no statistically significant main effects were found for the sample overall. Further, subgroup analysis revealed that DDMI resulted in significantly better treatment outcomes for the subgroup whose primary drug was cocaine, whereas SI resulted in significantly better treatment outcomes for the subgroup whose primary drug was marijuana.

Data for this pilot study were reported in Table 1 of Martino et al. (2006). We used the reported primary outcome *mean days of alcohol use for the past 28 days* to illustrate assurance as an estimate of sample sizes needed to sufficiently power a larger study. Interested readers can use our R program for other outcomes in Martino et al.

Table 1

Bayesian Assurance and Statistical Power Calculations Under Different Scenarios for Different Sample Sizes for Normally Distributed Data

Sample Size	Assurance	Power	Power 1	Power 2	Power 3	Power 4
20	0.260	0.186	0.086	0.336	0.241	0.126
40	0.397	0.333	0.134	0.596	0.438	0.215
60	0.481	0.467	0.181	0.773	0.602	0.301
80	0.537	0.585	0.227	0.880	0.728	0.384
100	0.578	0.682	0.273	0.939	0.819	0.461
120	0.609	0.760	0.318	0.970	0.883	0.533
140	0.633	0.822	0.362	0.986	0.926	0.598
160	0.652	0.869	0.405	0.994	0.954	0.656
180	0.669	0.905	0.446	0.997	0.971	0.707
200	0.681	0.932	0.485	0.999	0.983	0.752

Note. Assurance = the Bayesian assurance calculated by Monte Carlo simulation; Power = the calculated statistical power with the observed data; Power 1 = the calculated statistical power with the observed mean difference decreased by 1 day; Power 2 = the calculated statistical power with the observed mean difference increased by 1 day; Power 3 = the calculated statistical power with standard deviation decreased by 1 day; Power 4 = the calculated statistical power with standard deviation increased by 1 day.

(2006). For the whole sample ($N = 44$), the mean days for alcohol use for the past 28 days was 4.07 ($SD = 6.56$). With respect to the two study arms (i.e., DDMI vs. SI), the mean days for DDMI was 3.04 ($SD = 5.82$), and for SI the mean was 5.30 ($SD = 7.31$). Therefore, the mean difference between DDMI and SI was 2.26 days. The pooled *SD* can be calculated as 6.536, which would give the effect size of 0.346. The associated *t*-statistic is 1.142 ($df = 42$), which yields a one-sided *p* value of 0.130 indicating that DDMI is not significantly better than SI. This finding is consistent with the conclusions from Martino et al. (2006).

According to Cohen (1988), an effect size of 0.346 is a small to medium effect. If we used this effect size to design a future study to refine and further test DDMI, the sample size would need to increase to 133 participants per arm (i.e., 266 participants total). This is the sample size generated from a classical power analysis with a Type I error controlled at 2.5% (one-sided) with power of 0.80. With a population *SD* of 6.536, the sample size is based on a population mean difference of 2.26, which is the difference in days of alcohol use between the treatment and control conditions.

Using the sample size of 133 per arm, we ran the MC-Normal procedure 1 million times to estimate Bayesian assurance. So derived, assurance is 0.626 in comparison with the classical statistical power of 0.80 as designed. In other words, a sample size of 133 per arm produces power at 0.80 only if an effect size of 0.346 or greater is observed. Conditioned on all possible effect sizes (i.e., conditioned on greater uncertainty), the Bayesian perspective suggests that power with a sample of 133 participants per arm is only 0.626. This difference is not a surprise because Bayesian assurance is calculated as an average of classical statistical power over all possible prior specifications of mean differences and standard deviations (cf. Equation 2).

To further illustrate Bayesian assurance and classical statistical power, we ran the MC-Normal for sample sizes ranging from 20 to 200 participants and calculated the Bayesian assurance and the statistical power with the observed values from Martino et al. (2006). The findings are summarized in Table 1 and graphically illustrated in Figure 1. As seen in Table 1, Bayesian assurance is labeled as *Assurance* and statistical power with the observed values is labeled as *Power*.

In addition, we have included four supplemental scenarios as sensitivity analyses. These scenarios demonstrate the relationship between traditional statistical power estimates under different specifications of mean differences and their associated *SDs*. Scenarios 1 and 2 are designed to investigate sensitivity for mean differences with *SD*s fixed at the observed value. Labeled as Power 1, Scenario 1 decreases the observed mean treatment difference by 1 unit, from 2.26 to 1.26, which corresponds to a decrease in the effect size from the observed 0.346 to 0.193. Labeled as Power 2, Scenario 2 increases the observed mean treatment difference from 2.26 to 3.26, which corresponds to an increase in the effect size from the observed 0.346 to

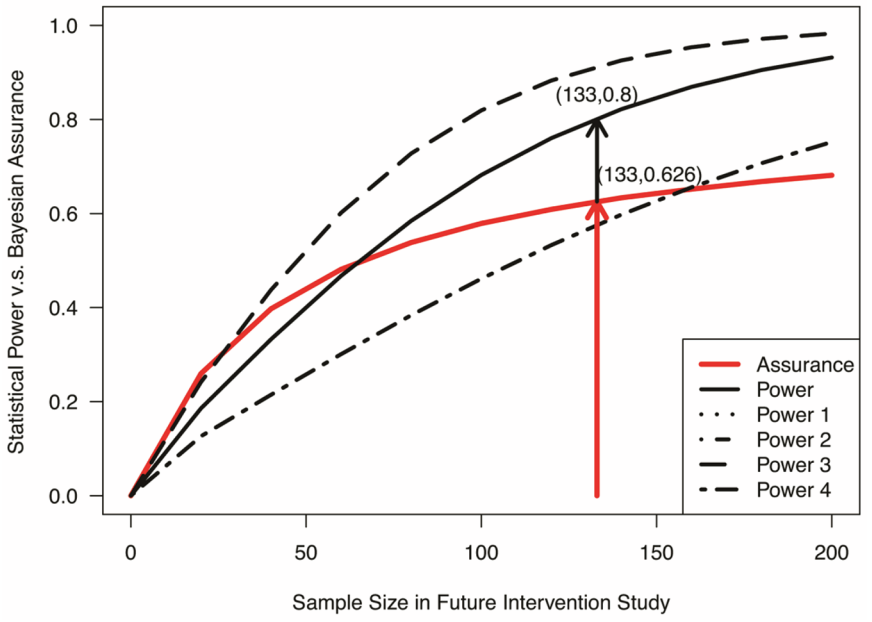

Figure 1. Calculated statistical power (Power; solid black line) and Bayesian assurance (Assurance; red line) with the observed mean difference and standard deviation for different specifications of sample sizes in a hypothetical future intervention study with continuous data. Power 1, Power 2, Power 3, and Power 4 are the statistical powers corresponding to the four scenarios of sensitivity runs reported in Table 1. The vertical arrows indicate a sample size of 133 participants where Bayesian assurance is 0.626 (red arrow) and statistical power is 0.8 (black arrow).

0.499. Similarly, Scenarios 3 and 4 are designed to investigate sensitivity related to *SD*s with the mean difference fixed at the observed value. Specifically, Scenario 3 (labeled as Power 3) decreases the observed *SD* by 1 unit, from 6.536 to 5.536, which corresponds to an increase in the effect size from the observed 0.346 to 0.408. Labeled as Power 4, Scenario 4 increases the observed *SD* by 2 units, from 6.536 to 8.536, which corresponds to a decrease in the effect size from the observed 0.346 to 0.265. Figure 1 illustrates the calculated statistical power and Bayesian assurance with the observed mean difference and *SD* for different specifications of sample sizes in a hypothetical future intervention study with continuous data.

In Table 1 and Figure 1, Bayesian assurance across the scenarios increases as sample sizes increase. Further, Bayesian assurance is generally smaller than statistical power for relatively large sample sizes (e.g., for motivational interviewing studies with sample sizes greater 60 participants per arm). However, Bayesian assurance is larger than classical power (column 3 in Table 1) in sample sizes that include fewer than 60 participants per arm. These larger values are expected because

assurance is the average effect for all possible specifications of distributions of the mean difference and the *SD*.

Also shown in Table 1 and Figure 1, statistical power is sensitive to different specifications of mean treatment differences and *SD*s. For example, for the impact of the mean treatment difference as seen in Power 1 and Power 2, by decreasing or increasing the mean treatment difference by 1 day on *alcohol use*, statistical power changes from 0.18 to 0.77 for a sample size of 60, from 0.22 to 0.88 for a sample size of 80, from 0.36 to 0.99 for a sample size of 140, and from 0.40 to 0.99 for a sample size of 160. Similarly, for the influence of *SD*s as seen in Power 3 and Power 4, by decreasing the observed *SD* from 6.536 by 1 unit in Power 3 and increasing the observed *SD* from 6.536 by 2 units in Power 4, the resultant statistical power could be dropped from 0.77 to 0.30 for a sample size of 60, from 0.879 to 0.394 for a sample size of 80, from 0.986 to 0.598 for a sample size of 140, and from 0.993 to 0.657 for a sample size of 160.

Randomized Trial of Specialty Mental Health Probation

As an example of calculating assurance with a binary dependent measure, we used data from a recently completed feasibility study as reported in Cuddeback (2016). For this study, individuals with serious mental illnesses (e.g., schizophrenia, bipolar disorder) who were on probation in two counties—one urban and one rural—in a large southeastern state were randomly assigned to receive either standard probation or a specialty mental health probation (SMHP). Briefly, the core components of SMHP included (a) a reduced probation caseload size, (b) an exclusively mentally ill caseload, (c) a problem-solving supervision orientation, (d) ongoing mental health training, and (e) greater connection to behavioral health and other community-based resources. Criminal justice and mental health outcomes were collected at baseline, 6-month, and 12-month intervals.

For this article, the probability of a technical violation (a binary outcome) was used to illustrate the calculation of assurance for binary data. Probationers are usually given a technical violation when they fail to meet the specific requirements of their probation. In the pilot study, 46 participants were randomly assigned to SMHP, and 50 were assigned to receive standard probation. Treatment and control subjects were balanced on all observed baseline covariates. Study findings suggested that 62% ($n = 31$) of the subjects who received standard probation had a technical violation, and 50% ($n = 23$) of those who received SMHP had a technical violation, which corresponds to a 12% decrease among experimental subjects. Using these values to design a future study with 160 participants per arm, statistical power would be 70% and the Bayesian assurance from MC-Binary would be 65% or slightly smaller.

To further illustrate the MC-Binary approach for other specification of sample sizes, we ran the Monte Carlo procedure for sample sizes from 50 to 300 partici-

pants. The resulting statistical power and Bayesian assurance with the observed values are labeled as *Statistical Power* and *Bayesian Assurance* in Table 2.

As with the continuous data, four additional analyses are included to demonstrate the sensitivity of traditional statistical power to different specifications of the number of participants who responded to the treatments. Scenarios 1 and 2 are designed to investigate the sensitivity for the responses from the SMHP treatment arm. Specifically, Scenario 1 (labeled as Power 1) decreases the observed response from 23 participants to 20. Scenario 2 (labeled as Power 2) increases the observed response from 23 to 26. Similarly, Scenarios 3 and 4 are designed to investigate the sensitivity for the responses from the standard probation arm. Specifically, Scenario 3 (labeled as Power 3) decreases the observed responses by 2 (from 31 to 29), and Scenario 4 (labeled as Power 4) increases the observed responses from 31 to 33. Figure 2 is an illustration of Table 2 depicting the calculated statistical power and Bayesian assurance with the observed response rates for different specifications of sample sizes in a hypothetical future intervention study with binary data.

Similar to the continuous data, Bayesian assurance increases as the sample size increases. Statistical power at the observed value (i.e., Power in column 3 of Table 2) is generally larger than the Bayesian assurance for a large sample size (> 100 in this study) but smaller for a small sample size (< 100). For the sample size 200 participants per arm, statistical power is about 0.80, whereas Bayesian assurance (i.e.,

Table 2
Bayesian Assurance and Statistical Power Calculations Under Different Scenarios for Different Sample Sizes for Binary Data

Sample Size	Assurance	Power	Power 1	Power 2	Power 3	Power 4
50	0.388	0.330	0.585	0.138	0.199	0.490
100	0.525	0.526	0.840	0.196	0.305	0.744
150	0.630	0.674	0.945	0.248	0.399	0.881
200	0.713	0.782	0.982	0.298	0.484	0.947
250	0.777	0.857	0.995	0.345	0.560	0.978
300	0.828	0.908	0.998	0.390	0.626	0.991

Note. Assurance = the Bayesian assurance calculated by Monte Carlo simulation; Power = the calculated statistical power with the observed data; Power 1 = the calculated statistical power with the response from specialty mental health probation (SMHP) decreased from the observed 23 participants to 20; Power 2 = the calculated statistical power with the response from SMHP decreased from the observed 23 participants to 26; Power 3 = the calculated statistical power with the response from the standard probation arm decreased from the observed 31 participants to 29; Power 4 = the calculated statistical power with the response from the standard probation arm increased from the observed 31 participants to 33.

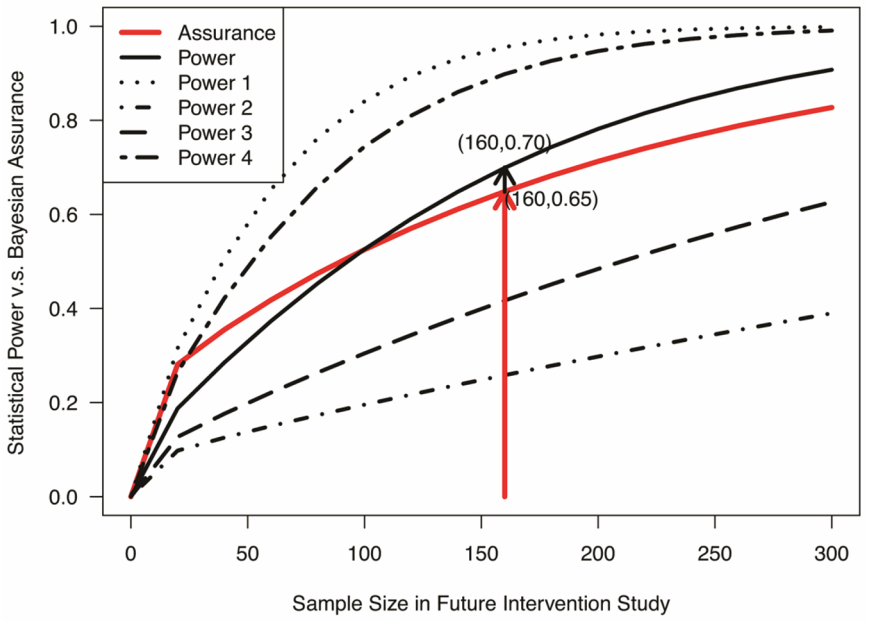

Figure 2. Calculated statistical power (Power; solid black line) and Bayesian assurance (Assurance; red line) with the observed response rates for different specifications of sample sizes in a future intervention study with binary data. Power 1, Power 2, Power 3, and Power 4 are the statistical powers corresponding to the four scenarios of sensitivity analysis in Table 2. The vertical arrows indicate a sample size of 160 participants where the Bayesian assurance is 0.65 (red arrow) and the statistical power is 0.70 (black arrow).

the expected probability of identifying a successful SMHP to reduce technical violation over standard probation) is about 71%.

Likewise, as seen in Table 2 and Figure 2, statistical power is sensitive to different specifications of the number of responses. Specifically, for the impact from the SMHP treatment condition as seen in Power 1 and Power 2, by decreasing or increasing the violation response by three participants, power could change from 0.58 to 0.14 for a sample size of 50, from 0.945 to 0.248 for a sample size of 150, and from 0.995 to 0.345 for a sample size of 250. Similarly, for the influence from the control condition as seen in Power 3 and Power 4, by decreasing or increasing the violation response by two participants, the resultant statistical power could be changed from 0.199 to 0.490 for a sample size of 50, from 0.399 to 0.881 for a sample size of 150, and from 0.560 to 0.978 for a sample size of 250.

Discussion

In this article, we introduced Bayesian assurance as an alternative to traditional power analysis in intervention research. We drew, in part, from work on clinical

trials (Du & Wang, 2016; O'Hagan & Stevens, 2001; O'Hagan et al., 2005; Ren & Oakley, 2014). In conventional research design, statistical power is calculated as the probability of rejecting the null hypothesis conditional on a specified intervention effect. Under assurance, power is calculated as the unconditional probability of a successful intervention averaged over the parameter space of a prior treatment effect and other nuisance parameters. The calculation of Bayesian assurance involves a high-dimensional integration that is computationally challenging. However, Monte Carlo simulation provides an adequate alternative estimation procedure. The Monte Carlo simulation-based approach is illustrated in R with continuous and binary data.

Bayesian assurance has an important advantage over traditional power estimation in intervention research. A traditionally powered intervention at 80% does not guarantee an 80% probability of success because the power calculation is based on a single pre-specified, fixed treatment effect that might be different from the true unobserved treatment effect. We demonstrated the sensitivity of traditional statistical power in Tables 1 and 2 by making slight changes in the observed values (i.e., mean differences and *SD*s) in continuous data and the response rates in binary data. The findings suggest that statistical power is sensitive to minimal misspecification, and misspecification could easily lead to false confidence derived, in part, from erroneous power estimates. The sensitivity of traditional power calculations to variation in expected effect sizes and *SD*s could contribute to explanations for why seemingly appropriately powered studies with well-conceived and well-implemented interventions sometimes produce nonsignificant and inconclusive findings.

An emerging alternative to the traditional approach, Bayesian assurance can be used for sample-size determination in the same way that classical statistical power is used. The sample sizes determined from Bayesian assurance can be easily obtained from Tables 1 and 2 (or in Figures 1 and 2). For example, using Table 1 or Figure 1, the sample size for the first study would be 120 per intervention arm if a researcher would like to have 60% assurance. Similarly, the sample size for the second study would be 200 per arm if the researcher would like to have 71% assurance. Cutoffs—60% for the first study and 71% for the second—are entirely determined by the researcher and substantive issues (e.g., the degree of risk a funder is willing undertake).

Typically, assurance is lower than conventional power because assurance is estimated across all possible effect sizes. Calculated in this manner, assurance usually suggests that larger samples will be needed. Because it considers a wide range of effect sizes, assurance is more robust against inaccurate prior information.

Notwithstanding, Bayesian assurance is dependent on the goodness of the prior distribution of effect sizes and other related design parameters. If the prior distribution is misspecified, assurance can be incorrect. However, in this situation classical statistical power could produce worse estimates because it is calculated from a single value from the misspecified distribution. In this sense, Bayesian assurance, as result of averaging the range of prior values, provides a more leveraged protection

against poor prior information. In high-stakes research, when the costs of making an erroneous decision based on findings from an underpowered study are great, assurance may be a preferred method for estimating sample sizes.

Author Notes

Ding-Geng Chen, PhD, is the Wallace H. Kuralt Distinguished Professor and Director of Statistical Development and Consultation at the University of North Carolina at Chapel Hill.

Mark W. Fraser, PhD, is the John A. Tate Distinguished Professor for Children in Need in the School of Social Work at the University of North Carolina at Chapel Hill.

Gary Cuddeback, PhD, is an associate professor in the School of Social Work at the University of North Carolina at Chapel Hill.

Correspondence regarding this article should be directed to Ding-Geng Chen via e-mail dinchen@email.unc.edu

Acknowledgments

We thank Ms. Diane C. Wyant for providing helpful comments on the draft of this paper.

References

Chen, D. G., & Chen, J. D. (Eds.). (2017). *Monte Carlo simulation-based statistical modeling*. ICSA Book Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-981-10-3307-0

Chen, D. G., & Ho, S. (2016). From statistical power to statistical assurance: It's time for a paradigm change in clinical trial design. *Communication in Statistics-Simulation and Computation*. Advance online publication. https://doi.org/10.1080/03610918.2016.1259476

Chen, D. G., & Peace, K. E. (2011). *Clinical trial data analysis using R*. CRC Biostatistics Series. Boca Raton, FL: Chapman & Hall.

Chen, D. G., & Peace, K. E. (2013). *Applied meta-analysis with R*. CRC Biostatistics Series. Boca Raton, FL: Chapman & Hall.

Chen, D. G., Peace, K. E., & Zhang, P. (2017). *Clinical trial data analysis using R and SAS*. CRC Biostatistics Series. Boca Raton, FL: Chapman & Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics*, *5*, 305–309. https://doi.org/10.1002/pst.232

Cuddeback, G. S. (2016). Statewide mental health training and specialty mental health probation: Final report submitted to the North Carolina Governor's Crime Commission and the North Carolina Department of Public Safety. Chapel Hill, NC: School of Social Work, University of North Carolina at Chapel Hill.

Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, *51*, 589–605. http://dx.doi.org/10.1080/00273171.2016.1191324

Fraser, M. W., & Galinsky, M. J. (2010). Steps in intervention research: Designing and developing social programs. *Research on Social Work Practice*, *20*, 459–466. https://doi.org/10.1177/1049731509358424

Fraser, M. W., Richman, J. M., Galinsky, M. J., & Day, S. H. (2009). *Intervention research: Developing social programs*. New York, NY: Oxford University Press.

Kimko, H. C., & Peck, C. C. (Eds.). (2010). *Clinical trial simulations: Applications and trends*. Advances in the Pharmaceutical Sciences Series. New York, NY: Springer.

Martino, S., Carroll, K. M., Nich, C., & Rounsaville, B. J. (2006). A randomized controlled pilot study of motivational interviewing for patients with psychotic and drug use disorders. *Addiction*, *101*, 1479–1492. https://doi.org/10.1111/j.1360-0443.2006.01554.x

O'Hagan, A., & Stevens J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, *21*, 219–230. https://doi.org/10.1177/02729890122062514

O'Hagan, A., Stevens, J. W., & Campbell, M. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, *4*, 187–201. https://doi.org/10.1002/pst.175

Ren, S., & Oakley, J. E. (2014). Assurance calculations for planning clinical trials with time-to-event outcomes. *Statistics in Medicine*, *33*, 31–45. https://doi.org/10.1002/sim.5916

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004) *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: Wiley.

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. https://doi.org/10.1002/sim.4780050103

Manuscript submitted: May 9, 2017
Revision submitted: June 13, 2017
Accepted: June 19, 2017
Electronically published: Month XX, 2018