HHS Public Access

Author manuscript

J Biopharm Stat. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

J Biopharm Stat. 2018 ; 28(5): 857–869. doi:10.1080/10543406.2017.1399898.

Sample Size Calculations for Blinding Assessment

Victoria Landsman[1,2], **Mark Fillery**[3], **Howard Vernon**[3], and **Heejung Bang**[4]

[1]Institute for Work and Health

[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[3]Canadian Memorial Chiropractic College, Toronto, Ontario, Canada

[4]Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, California, USA

Abstract

Blinding is a critical component in randomized clinical trials along with treatment effect estimation and comparisons between the treatments. Various methods have been proposed for the statistical analyses of blinding-related data but there is little guidance for determining the sample size for this type of data, especially if blinding assessment is done in pilot studies. In this paper, we try to fill this gap and provide simple methods to address sample size calculations for a 'new' study with different research questions and scenarios. The proposed methods are framed in terms of estimation/precision or statistical testing to allow investigators to choose the best suited method for their goals. We illustrate the methods using worked examples with real data.

Keywords

blinding index; clinical trial; contingency table; masking; patient blinding

1. Introduction

Blinding has been widely perceived to be important in randomized controlled trials (RCT) and other comparative evaluations. Traditionally, blinding-related issues have been discussed more qualitatively or conceptually, for example, "blinding is important" and "double (or triple) blind is the best" (Hopton and Macpherson, 2011; Jadad, et al., 1996; Kolahi, et al., 2009; Wilsey, et al., 2016). In the last two decades, a growing number of studies on blinding have pursued a quantitative approach to study design, data collection, analysis and interpretation (Arandjelović, 2012; Bang, et al., 2010; Chow and Shao, 2004; James, et al., 1996; Jeong, et al., 2013; Wilsey, et al., 2016; Wright, et al., 2012). Nowadays, blinding is also emphasized in non-pharmacological trials, like trials involving devices or physical therapy, in order to demonstrate internal validity. Meta-analyses of blinding have offered some optimistic news about the feasibility of blinding for some interventions, which

Address correspondence to: Heejung Bang, PhD, One Shields Avenue, Med Sci 1C, Davis, CA 95616, USA. Phone: 530-752-6287; hbang@ucdavis.edu.

Conflict of Interest: None

are traditionally believed to be hard to blind, including non-drug or non-injection (Boutron, et al., 2007; Brinjikji, et al., 2010; Freed, et al., 2014; Hopton and Macpherson, 2011; Houweling, et al., 2014; Moroz, et al., 2013; Wilsey, et al., 2016).

Despite various statistical and methodological proposals, there is still little guidance concerning the determination of minimal or adequate sample size (N) for blinding assessment in a statistically justifiable manner. In the past, it was not unusual for a statistician to advise clients seeking sample size calculations for a blinding assessment to take a sample of at least N=30, or perhaps N=100 patients, in the absence of a good reference. In some cases, statisticians may have even said, "Use any N available". Recently, Shin et al. (Shin, et al., 2016) proposed a pilot RCT on acupuncture with 2 centers and selected a blinding index (BI) as one of the study outcomes (Bang, et al., 2004). This team recruited 40 participants which may be reasonable for a pilot study testing a clinical outcome, but it is unclear whether it is sufficient for testing a blinding outcome. In another example, Vernon and his colleagues proposed a new study for blinding assessment of real vs. control cervical manipulation procedures with multiple chiropractors following an initial study (Vernon, et al., 2013). The investigators decided to use a BI for evaluation of blinding as the primary outcome, but found a lack of statistically sound recommendations in the blinding literature to inform the choice of the sample size (Vernon, 2017) https://clinicaltrials.gov/ct2/show/NCT01772966.

The growing interest in statistical analyses of blinding outcomes in recent years (Arandjelović, 2012; Baethge, et al., 2013; Crisp, 2015; Hertzberg, et al., 2008; Houweling, et al., 2014; Wright, et al., 2012) creates the need to address sample size calculations in order to improve the quality of inference on analysis of blinding data. The key question is how to determine the sample size for a new stand-alone study, such as a pilot study focusing on masking, evaluation of short-term blinding (Walter, et al., 2005), or part of a large phase III RCT, in which blinding assessment may be defined as a secondary aim.

In this article, we propose simple and intuitive ways to address this question. Since blinding studies, unlike evaluations of clinical effectiveness, do not have clear-cut aims, we present three different research scenarios and describe methodologies to obtain sample sizes in each case. The proposed methods can be integrated into traditional frameworks (e.g., estimation or statistical testing) with clear underlying mechanisms and operational characteristics, and the calculations are straightforward. Of note, we recommend the proposed methods for power calculations while designing a 'new' study, and that post-hoc power calculations be avoided (CONSORT, 2010; Hoenig and Heisey, 2001).

2. Methods: background, notation, and proposals

Studies focusing on treatment comparison or survey research generally have clear and well-defined goals, such as detection of treatment effect with high power, or ensuring the margin of error not exceeding a desired bound. Conversely, blinding assessment studies tend to have somewhat subjective or varied goals. Regardless of the divergent goals, blinding studies are driven by two questions at the design and analysis stages: 1) was the blinding broken? and 2) if so, are the outcome data affected? These two questions can be converted into statistically

relevant scenarios reliant on the format of blinding assessment data. Typically, blinding data are presented as a 2×3 table with two allocation arms (Treatment (T=1) vs. Control (C=2)) and three choices for guess (1(=T), 2(=C), 3(=Don't know)) as in Table 1. Some researchers use a 2×2 format without the 'Don't know' option, while some others use a 2×5 format accounting for degree of belief (e.g., 'strongly believe', 'somewhat believe'), which can be reduced to a 2×3 format (Bang, et al., 2004; James, et al., 1996; Mathieu, et al., 2014; Wright, et al., 2012).

In this view, the three statistically relevant scenarios would be: 1) testing the independence of allocation and guess; 2) estimating arm-specific trinomial proportions of guess and their contrast; and 3) testing the effect of allocation-guess interaction on the clinical outcome. In the rest of this section, we derive the formulas for the sample size required for each of the three scenarios with adoption or adaptation of existing methods or ideas, followed by the Worked examples in the next section.

Let n_{ij} represent a cell count in a 2×3 table with blinding data (see Table 1; i=1,2 and j=1,2,3). The row sums, $n_1.$ and $n_2.$, define the sample sizes for each arm, and the total sample size is equal to $N=n_1.+n_2.$ Unless specified differently, we assume 1:1 allocation ($n_1.=n_2.=n$ and $N=2*n$) and $\alpha=0.05$. Also, we denote joint and conditional probabilities as p_{ij} and $p_{j|i}$, respectively. Parameter and estimator notation may be used interchangeably when the context is clear, as it is a common practice in research on N/power.

2.1 Scenario 1: Testing the independence of allocation and guess

The first and most natural consideration for an investigator seeing data in a 2×3 table (as in Table 1) would be a classical test of the independence of allocation (row) and guess (column) or the homogeneity of the response in multiple groups (Agresti, 2013; Mathieu, et al., 2014). The Pearson Chi-square and Likelihood Ratio (LR) tests are standard tests for the association in an unordered rxc table, with r arms and c guesses.

Under the alternative hypothesis, Chi-square and LR statistics, commonly denoted as X^2 and G^2 in the literature, have large-sample noncentral chi-squared distributions. Let p_{ij} denote the joint probability in cell (i,j), where i represents the allocation (i=1,2), j represents the guess (j=1,2,3), and $\bar{p}_{ij}$ denote the joint probability in cell (i,j) under the null (the independence hypothesis); $\sum p_{ij} = \sum \bar{p}_{ij} = 1$. Using this notation, the noncentrality parameter (λ) for Pearson Chi-square equals

$$\lambda = N \sum_{i,j} \left(p_{ij} - \bar{p}_{ij}\right)^2 / \bar{p}_{ij},$$

and the noncentrality parameter for LR statistic equals

$$\lambda = 2N \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{\bar{p}_{ij}}\right).$$

The desired sample size (N) for a chi-squared test can be obtained from the power equation $P\left(X^2_{\upsilon,\lambda} > \chi^2_{\upsilon}(\alpha)\right)$, where $\upsilon = (r-1)\times(c-1)$ is the degrees of freedom (df). If the data from 'previous' studies are available and we want to detect the similar observed differences, p_{ij} and $\bar{p}_{ij}$ can be estimated as n_{ij}/N and $(n_{i}.*n_{.j})/N^2$, respectively. Approximations can be done using published or built-in table, e.g., (Agresti, 2013); see the Worked Examples below.

2.2 Scenario 2: Estimation of (a) arm-specific trinomial proportions and (b) their contrast

Blinding data for a given arm i can be viewed as a sample from a trinomial distribution with probabilities $\left(p_{1|i}, p_{2|i}, p_{3|i}\right)$, and $\sum_{j=1}^{3} p_{j|i}=1$. In this subsection, we consider determination of the sample size required in each arm (n) to ensure the estimation of these probabilities and/or their difference (namely, BI) with a specified level of precision.

First, let us consider the estimation of the probabilities. In this case, the objective is to find the smallest sample size n for arm i such that the estimated proportions are *simultaneously* within specified distances of true population proportions with a probability of at least 1-α, that is,

$$P\left(\bigcap_{j=1}^{3}\left|\hat{p}_{j|i} - p_{j|i}\right| \le d_j\right) \ge 1-\alpha \text{for arm } i\,.$$

The sample size determination procedures proposed by Tortora (Tortora, 1978) and Thompson (Thompson, 1987) are based on the simultaneous confidence intervals (CIs) for the multinomial model. Tortora constrained the width of the jth category interval, j=1,2,3, to be $\leq d_j$ and obtained the following formula for n,

$$n = \max_{j=1,2,3}\left[z^2_{\alpha/(2\times 3)} p_{j|i}\left(1-p_{j|i}\right)/d_j^{\,2}\right]$$

where $z_{\alpha/(2\times 3)}$ is the upper (α/6)*100th percentile of the standard normal distribution. The total sample size N can be obtained as N=2*n, where n is the larger value between the sample sizes obtained for each arm. In the absence of prior knowledge about $p_{j|i}$, the 'worst-case' (in view of the maximal n required) would be to use the probabilities vector (1/2,1/2,0) for each arm. Assuming $d_j = d$ for each category j, the (maximal) n for each arm can be further simplified as

$$n = \frac{z^2_{\alpha/6}}{4d^2} = \frac{1.43279}{d^2} \text{for } \alpha = 0.05\,.$$

In contrast, Thompson described the general form of the 'worst' parameter vector under the constraint of the equal width, $d_j = d$. His theoretical result depends on the specified level of

α: for example, (1/3,1/3,1/3) is the 'worst-case' parameter vector if α=0.05, whereas (1/2,1/2,0) is the 'worst-case' parameter vector if α=0.025; see Table 1 in (Thompson, 1987). His procedure results in smaller sample sizes and hence is attractive under the condition of the equal width intervals. The conservative sample size using the Thompson's method is given by

$$n = \frac{2 \cdot z^2_{\alpha/6}}{9d^2} = \frac{1.27359}{d^2} \text{ for } \alpha = 0.05 \quad \text{Eq (1)}$$

which is lower than the Tortora's counterpart. Both methods would result in sample size estimates that ensure a specified confidence range for the estimated probabilities.

Next, we consider the estimation of the BI. Bl is an arm-specific index for blinding assessment defined as a contrast between the probability of correct guess and the probability of incorrect guess:

$$\mathrm{BI_T} = \mathrm{p}_{1|1} - \mathrm{p}_{2|1} \text{and } \mathrm{BI_C} = \mathrm{p}_{2|2} - \mathrm{p}_{1|2}.$$

In a 2×2 format without the 'Don't know' option, the BI reduces to $\mathrm{BI_i} = 2\mathrm{p_i} - 1$ for $i = T, C$, where $p_T = p_{1|1}$ and $p_C = p_{2|2}$. Estimators for the BIs can be obtained by replacing the arm-specific probabilities by their estimators:

$$\widehat{\mathrm{BI}}_\mathrm{T} = \left(n_{1T} - n_{2T}\right)/n_T \, and \widehat{\mathrm{BI}}_\mathrm{C} = \left(n_{2c} - n_{1c}\right)/n_c.$$

BI quantifies the correct guess beyond chance (50%) or imbalance between correct vs. incorrect guesses in the blinding data as in Table 1. For example, BI=0 represents 'random guess'; BI=0.3 represents a 30% imbalance in guess toward the correct guess, say, 40% participants guessed T vs. 10% guessed C in arm T; BI=−0.3 represents an imbalance of 30% toward the incorrect guess. BI is usually interpreted in terms of possible blinding scenarios, along with qualitative data (e.g., reasons for guess), whenever available. After all, there are contexts in which correct guesses are not undesirable, and may provide insight, such as when they reflect 'wishful thinking' (Bang, 2016; Brinjikji, et al., 2010).

A standard 2-sided CI for BI with (1-α) confidence level for the treatment arm i is:

$$(\hat{p}_{1|i} - \hat{p}_{2|i}) \pm z_{\alpha/2} \cdot \sqrt{\left[\hat{p}_{1|i}\left(1 - \hat{p}_{1|i}\right) + \hat{p}_{2|i}\left(1 - \hat{p}_{2|i}\right) + 2\hat{p}_{1|i}\hat{p}_{2|i}\right]/n_i}.$$

Assuming 1:1 allocation as before, the objective is to find a sample size n such that the inequality $z_{\alpha/2} \cdot \sqrt{[\hat{p}_{1|i}\left(1 - \hat{p}_{1|i}\right) + \hat{p}_{2|i}\left(1 - \hat{p}_{2|i}\right) + 2\hat{p}_{1|i}\hat{p}_{2|i}]/n} \leq d$ holds for a specified threshold d. Solving this inequality in terms of n yields

$$n = z^2_{\alpha/2}\left[\hat{p}_{1|i}\left(1-\hat{p}_{1|i}\right)+\hat{p}_{2|i}\left(1-\hat{p}_{2|i}\right)+2\hat{p}_{1|i}\hat{p}_{2|i}\right]/d^2.$$

Using the method of Lagrange multipliers, we can show that the maximal sample size is reached when $p_{3|i}=0$ and $p_{1|i}=p_{2|i}=(1-p_{3|i})/2$. In this case, the maximal sample size is attained at the trinomial vector $(p_{1|i}=p_{2|i}=p_{3|i})=(1/2, 1/2, 0)$ simplifying the above formula to $n=z^2_{\alpha/2}/d^2$. This result has two important implications: a) allowing 'Don't know' as a guess category decreases the required sample size; and b) in a 2×2 format, the maximal value of $\mathrm{Var}(\hat{BI}_i)= 4*\mathrm{Var}(\hat{p}_i)$ is reached at the binomial vector (1/2, 1/2), resulting in the maximum sample size equal to $n=z^2_{\alpha/2}/d^2$. Interestingly, this sample size is 4 times of the well-known conservative sample size for estimation of the probability of event in binomial data.

To summarize, with good estimates for trinomial or binomial probabilities, the sample size for each arm is given as $n=z^2_{\alpha/2}[p_{1|i}(1-p_{1|i})+p_{2|i}(1-p_{2|i})+2p_{1|i}p_{2|i})]/d^2$ for a 2×3 format, and $n = z^2_{\alpha/2}[4p_i(1-p_i)]/d^2$ for a 2×2 format. In the absence of good estimates, the conservative sample size

$$n = z^2_{\alpha/2}/d^2 \quad \text{Eq (2)}$$

with p=1/2 can be used for both formats. It can be seen from these formulae that a fundamental operational characteristic, common in virtually all sample size estimations, applies here as well: the more stringent the threshold (or the narrower the CI), the larger the sample size required.

2.3 Scenario 3: Testing the effect of allocation-guess interaction on the clinical outcome

Breached blinding has a potential to distort clinical findings, leading to biased estimates of treatment effects with unknown direction (Bang, 2016; Mathieu, et al., 2014). In the presence of allocation-guess interaction, the estimated average treatment effect (ATE) may depend on the guess status (1(=T), 2(=C), 3(=Don't know)) resulting in meaningfully different ATEs in subgroups of different guess status. For instance, the ATE estimate obtained from those who guessed T can be positive, in favor of treatment, while the estimate from those who guessed C can be negative or null. In these situations, detecting the interaction between allocation and guess with a reasonable power could be of scientific interest.

Testing the effect of allocation-guess interaction on a (continuous) clinical outcome may be understood in a framework of a two-way fixed effects unbalanced ANOVA. Let y_{ijk} be the outcome from the kth patient in arm i with guess j, where i=1,2; j=1,2,3; and k=1,…,n_{ij}, and the cell sizes n_{ij} (non-zero) are defined in Table 1. In a 2×3 case, a univariate general linear model can be parametrized as

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk},\ i = 1,2;\ j = 1,2,3$$

where ε_{ijk} are normally distributed independent and identical errors with mean zero and variance σ^2. This model can be written in matrix notation as $y = XB + \varepsilon$, where X is a design matrix of size $N \times m$ of zeros and ones (m=6 in our case), and $B = \left[\mu_{11}\ \mu_{12}\ \mu_{13}\ \mu_{21}\ \mu_{22}\ \mu_{23}\right]$ (Elston and Bush, 1964).

Testing the null hypothesis of no interaction between allocation and guess is equivalent to testing the equality of ATE for each category of guess. This hypothesis can be viewed as a special case of a linear hypothesis H_0: LB=0 vs. H_A: LB ≠0, with L being a $q \times m$ contrast matrix of full rank, where q is the number of contrasts. For example, for an overall test of no interaction effect, $L = \begin{bmatrix} 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{bmatrix}$. The F-statistic is given by

$$F = \frac{\left(L\widehat{B}\right)'\left(L(X'X)^{-1}L'\right)^{-1}\left(L\widehat{B}\right)/q}{e'e/(N-m)}$$

with $\hat{B} = (X'X)^{-1}X'y$ and $e'e = \left(y - X\hat{B}\right)'\left(y - X\hat{B}\right)$. Under H_A, F is distributed as F(q, N-m, λ) with a noncentrality parameter $\lambda = (LB)'(L(X'X)^{-1}L')^{-1}(LB)/\sigma^2$. The sample size is computed by inverting the power equation P(F(q, N-m, λ) ≥F_α(q, N-m)); see (Castelloe and O'Brien, 2001; Elston and Bush, 1964; Muller and Peterson, 1984; O'Brien, 1986; O'Brien and Shieh, 1992) for details and general theory.

In summary, use of this method requires outcome data for allocation by guess status from historical studies, preliminary data, or pre-specified values (Chow and Shao, 2004; Wright, et al., 2012). Thus, the utility of this method may be limited in practice when it is difficult to come up with plausible inputs, although it may be of ultimate interest related to blinding assessment. See Section 3.3 for an illustrative example from Chow and Shao, where suspected breached blinding might have been problematic and warranted further investigation.

3. Worked examples

Even with technically sound methods, a crucial issue is how to implement the methods and make sensible decisions in practice. In this section, we illustrate sample size calculations using inputs from published data to assist in designing a 'new' study.

If a new study is designed as a pilot study where blinding is the primary outcome, trialists may use our N methods directly (e.g., a wide margin can be predefined for a pilot, and a narrow margin for a real study). If blinding is defined a secondary or tertiary outcome in a 'new' study, trialists may choose N for primary clinical outcome, which is typically done in RCTs (Briggs, 2000). If N was obtained for a primary clinical outcome, we can calculate what power or precision of the estimate of the blinding parameter can be attained with this N. Toward making an overall conclusion, we adopt Cohen's logic (Cohen, 1990): determine the sample size necessary to detect a negligible signal/indication of breached blinding with high probability. After the research is carried out using that sample size, and the result is not significant, the conclusion is justified that no nontrivial signal exists, at a given level. This

does, in fact, *probabilistically show* the intended null hypothesis of no more than a trivial small signal (i.e., blinding is acceptable).

3.1. Scenario 1: Testing the independence of allocation and guess

The sample size under this scenario can be obtained manually or using SAS macros powerRxC or Unifypow (SAS Institute, Cary, NC), among others (Castelloe and O'Brien, 2001; O'Brien and Shieh, 1992). As an illustrative example of a manual calculation, the real (rather than hypothetical) data from a study for blinding assessment of a sham cervical manipulation procedure (Vernon, et al., 2013) is used as an input for calculation of a sample size for a new study. Rigorous blinding evaluation was chosen as a primary aim of the new study (Vernon, 2017). The self-reported guess status collected from a secondary analysis in an earlier study is presented in Table 2 and it may be used to obtain the observed joint probabilities in a 2×3 table: 0.25, 0.14, 0.11 for arm T and 0.14, 0.23, 0.125 for arm C for guess 1,2,3, respectively. The expected joint probabilities under independence are: 0.195, 0.1875, 0.115 for guess of 1,2,3 in both arms. Using this information, the noncentrality parameter for Pearson Chi-square equals λ=0.054*N. The approximate power for a number of different N values can be obtained from the table 'Power of Chi-squared Test for α=0.05' in (Agresti, 2013) or various statistical software: if N=20, λ = 1.08, the power is approximately 0.13 (with df=2). Similar calculations show that N=176 will be required to test the same hypothesis with 80% power. In addition, the noncentrality parameter for the LR test equals λ=0.055*N, which is very similar to a Pearson's as expected. Sample SAS codes are provided in Appendix. Of note, resulting estimates can be unstable or unreliable with low cell counts.

3.2. Scenario 2: Estimation of (a) arm-specific trinomial proportions and (b) their contrast

Proportions obtained from previous studies can be used to inform sample size calculations. For example, Vernon et al. (Vernon, et al., 2013) observed the trinomial proportions (0.5, 0.28, 0.22) in arm T and (0.28, 0.47, 0.25) in arm C in an immediate post-treatment evaluation of blinding. Park et al.'s acupuncture trial (Park, et al., 2005) also observed the proportions of guess near 0.5 (e.g., 26/49). In the absence of prior data on trinomial proportions or with the observed proportions close to 0.5, as in the trials above, the conservative sample size for one arm can be obtained using $n = \frac{z^2_{\alpha/6}}{4d^2}$. For example, with α=0.05, the sample size equals n=574, 144, 36 for d=0.05, 0.1, 0.2, respectively. The sample sizes corresponding to the Thompson's method equal n=510, 128, 32. Since in the blinding context, $p_{1|i}$ and $p_{2|i}$ near 0.5 are quite plausible (whereas extreme scenarios such as (0;0;1) are rare), these conservative sample sizes are reasonably justified.

Next, we discuss the estimation of BI, the contrast of the arm-specific proportions. Recent systematic reviews and meta-analyses on blinding provide a new insight on the range of feasible values of the BI in different types of studies (Baethge, et al., 2013; Freed, et al., 2014; Moroz, et al., 2013). For example, Freed et al. (Freed, et al., 2014) focused on meta-analysis of BI in trials of psychiatric disorders. It is remarkable that a large number of studies included in Freed et al. produced the BI (in weighted average) close to 0 in the

control arm, where BI=0 corresponds to a 'random guess', supposedly the most ideal blinding scenario (Bang, 2016). Therefore, in the absence of good input about a possible value of BI in a future study, BI=0 could be a reasonable starting point for sample size calculation.

For a blinding data collected in a 2×3 format, BI=0 is implied by all parameter vectors of the form of $\left(p_{1|i}, p_{1|i}, 1-2p_{1|i}\right)$, where $0 \le p_{1|i} \le 1/2$. Table 3 presents sample calculations for this case obtained for two thresholds (d=0.1 and d=0.2) for $p_{1|i} = 0.1, 0.2, 0.3, 0,4, 0.5$. The results in the table clearly demonstrate that as $p_{1|i}$ values get closer to zero (implying that more people are expected to answer 'Don't know'), the required sample size decreases. When $p_{1|i}$ value gets closer to 0.5, a larger sample size is required. The 'worst-case' corresponds to $p_{1|i} = 0.5$, in which case 2×2 and 2×3 formats are equivalent and the required sample size reaches its maximum value. All confirm our theoretical results in Section 2.2. Notice a 5-fold difference in sample size required for the case (0.1, 0.1, 0.8) as opposed to the case (0.5, 0.5, 0) for the same value of α and d. The tighter the desired width of the CI (=2*d), the larger the sample size required, e.g., 4 times for d=0.1 vs. 0.2.

As an illustration, when a future study on blinding assessment of cervical manipulation is designed after a pilot study (Vernon, et al., 2013), and the team anticipates BI values to be in the range [0, 0.1], the required sample size per arm would be 86 (assuming parameter vector of (0.5, 0.4, 0.1)) or 97 (assuming (0.5, 0.5, 0)) with d=0.2. These sample sizes correspond to N=2*86=172 or N=2*97=194 which is comparable to the N=176 obtained under Scenario 1.

As noted above, our methods can also be used to assess the power or estimation precision for blinding (defined as a secondary or tertiary outcome) in the studies with the sample size obtained for clinical outcome. To exemplify the use of our methods for this common scenario, assume that a research team is planning a new trial to estimate the effect of real vs. sham acupuncture on muscle spasticity as primary outcome. The Ashworth scale for muscle spasticity is defined as a dichotomized outcome (yes/no increase in muscle tone). The team is informed by a previous trial (Park, et al., 2005) and wants to detect a clinically meaningful difference of 33% vs. 22% for real vs. sham acupuncture, respectively. The sample size n=276 for each arm is required to detect this difference with α=5% and 80% power with a 1:1 allocation via the Fisher exact test. If we decide to use the Thompson's formula for a trinomial vector and a conservative formula for BI, Equations (1) and (2), we get d=0.07 and 0.12, respectively, with the given n. Thus, we may expect that the total sample size N=552 is sufficient to make the goal of reliable blinding assessment achievable in the planning stage.

3.3. Scenario 3: Testing the effect of allocation-guess interaction on the clinical outcome

Chow and Shao (Chow and Shao, 2004) analyzed the Brownell and Stunkard's data (Brownell and Stunkard, 1982) focusing on breached blinding, the role of consent form, and its potential impact on the clinical outcome in the weight loss trial. The observed blinding data (Table 4) yielded substantial agreement between the allocation and guess and very high BI values (BI_T=0.67 and BI_C=0.52; markedly larger than the previously suggested threshold,

0.2 (Freed, et al., 2014; Kolahi, et al., 2009; Park, et al., 2008)), which may be indicative of possible breach in blinding. Moreover, dramatic interaction between allocation and guess on the clinical outcome of weight loss is apparent in Figure 1, created using the summary statistics provided in the original papers (Brownell and Stunkard, 1982; Chow and Shao, 2004). These summary statistics may be used to reproduce the raw outcome data closely.

In this and similar situations, we would be interested in the sample size required to test the overall interaction (i.e., the equality of ATEs for all guess categories) as well as the custom interaction (e.g., the equality of ATEs between those who guessed T and those who guessed C) with sufficient power in a new study. The method for sample size determination described in Section 2.3 can be implemented using GLMpower procedure in SAS; see Appendix for a sample code. Cell means, cell counts and error variance σ^2 are the sample required inputs. An estimate of the error variance may be obtained by fitting a two-way ANOVA model to the re-constructed raw data or crudely approximated as $\sigma \approx (y_{max} - y_{min})/6$, where y_{max} and y_{min} stand for maximal and minimal values of a reasonably symmetric outcome. Using the re-constructed raw data, we obtained ≈ 5, thus confirming the previous estimation (Chow and Shao, 2004). We ran an additional power analysis with $\sigma \approx (4 - (-20))/6 = 4$ using information provided in raw data plot (Brownell and Stunkard, 1982) for an illustration purpose.

Again, let us assume that we are planning to design a new trial, with the primary outcome defined as weight loss and blinding being selected as one of the secondary outcomes. We hypothesize in this case that the clinically meaningful difference in weight loss between treatment and placebo is about 3.5 kg with a conservative value σ=5. Assuming a 1:1 allocation, α=1% and power=95% (these strict conditions are assumed to minimize false positive results and ensure very high power), n=74 for each arm (total N=148) will be required to test the difference in means.

Using the results in Table 4, we can conclude that this sample size will be sufficient to detect overall and custom interactions defined above for blinding with high power (>90%) at α=5%. Let us now consider another, conservative situation with sample size N=98. This sample size is ~30% lower than N=148 but is double the original sample size (N=49) used in Brownell and Stunkard (1982). Using this sample size with the standard assumptions of α=5% and power=80%, we will be able to detect a true difference of <3 kg in treatment vs. control groups. Using the results in Table 4, we can now conclude that a sample size in the range of 100-150 would be sufficient to detect interactions of interest defined above with power >80% assuming $\sigma \approx 5$ and α=5%.

On the other side, the power to detect the difference in ATE among those who guessed T vs. those who answered 'Don't know' is very low (~5%), which is expected as the associated ATE lines have nearly the same slopes in Figure 1. At the same time, we might assume that those who chose 'Don't know' tend to be neutral or less biased. A similar analysis can be repeated with a different classification for guess status: guess correctly vs. guess incorrectly vs. not guess, as attempted by Chow and Shao.

This exercise demonstrates the feasibility of sample size and power calculations in the setting where we design a new study and reasonable blinding data along with the outcome data are available for inputs as educated guess. Although this type of data is rare in the present literature, we believe that growing research on blinding will yield more collection and reporting of similar data.

4. Discussion

In this paper, we consider the three qualitatively different scenarios relevant to quantitative analysis of blinding, and present methods of sample size determination for planning a future study with a blinding assessment component. The scenarios are framed in terms of estimation and precision (Scenario 2a & b)) as well as statistical testing and power (Scenario 1 and 3). We illustrated the three — real and hypothetical — scenarios with Worked examples using published data and exemplified the calculations manually or using a statistical software.

The methods and examples in this paper offer users a suite of formulae that can be used to determine sample sizes in order to conduct a spectrum of studies, from a pilot or feasibility study (Walter, et al., 2005) to a full scale RCT. Since power and sample size calculation should be tailored to a specific research question, the choice of the particular method and formula depends on the goals and input availability. The proposed methods could be particularly essential for studies testing and establishing a newly developed control, sham or placebo intervention. Blinding assessment is also desirable for studies that test if the two treatments are easily distinguishable and there are no clinical outcomes (say, masking).

The proposed methods can be implemented in a flexible manner for different purposes. For example, when designing a pilot study, researchers may decide to use the formulae in the paper with a wider margin of error than that which is suitable for a real trial: for example, 0.2 can be set as a targeted threshold for estimation of the BI in a pilot study, and 0.1 in a larger, actual trial. If a research team decides to evaluate blinding in a very large trial, the sample size formulae may be used to create a subsample from the entire sample, e.g., (COMMIT, 2005), to which a blinding questionnaire could be administered. Similarly to power/sample size analysis relevant to a clinical outcome, if the sample size used in blinding assessment is substantially lower than the ones justifiable by the methods in our paper, the designation of “pilot or feasibility” study may be reasonable.

We want to discuss some limitations. First, we considered studies with a 1:1 allocation and 2 arms since blinding data is optimally interpreted in this setting. With a 2:1 allocation, we are not fully certain if P(correct guess)=0.5 is still the most ideal value; some patients may use 50% in their guess (between T vs. C), while others may use 66.7% if they were informed about the allocation ratio and remember it (Bang, 2016; Brownell and Stunkard, 1982). N/power calculations for advanced designs (e.g., with 2:1 allocation, clustering or crossover designs) as well as accounting for other complex issues (e.g., informative drop-out) in RCT are possible topics for future research (Bang, et al., 2010; Park, et al., 2005; Roy, 2012; Zhang, et al., 2013).

Second, we did not handle multiple testing rigorously, partly because we do not pursue rigorous hypothesis testing in blinding. Of note, we framed the problem in classical superiority testing with a 2-sided CI/test, not equivalence test or a 1-sided hypothesis/CI, which may be more relevant to blinding. The main reason behind our decision was to avoid unnecessary complexity (e.g., equivalence margin), because blinding is a tool, not a goal, and any numerical analysis alone should not be used for binary designation (e.g., success or failure) (Bang and Park, 2013; Zhang, et al., 2013). Along the same line, we emphasize the importance of the estimation-based method described in Scenario 2 in the blinding context. Finally, post-hoc or retrospective power calculation should be avoided (CONSORT, 2010; Hoenig and Heisey, 2001).

In closing, we propose the methods for sample size and power calculations which can be used for exploratory or planning purposes, and can address different research questions, inputs and settings commonly encountered in blinding.

Acknowledgments

Funding: HB was partly supported by the National Institutes of Health through grants UL1 TR001860 and P50 AR063043. HV was supported by the National Institutes of Health through grant 1R21AT004396.

Appendix: Sample SAS codes

1. Scenario 1

```
data Vernon;
  do Treat=1 to 2;
    do Guess=1 to 3;
      input freq @@; output;
    end;
  end;
datalines;
16 9 7
9 15 8
;
%powerRxC(data=Vernon,row=Treat,col=Guess,count=freq,nrange=
%str(20,50,100,176,200 to 500 by 100))
```

2. Scenario 3

```
data Chow;
input Treat Guess Outcome Weight;
datalines;
1 1 9.6 19
1 2 3.9 3
1 3 12.2 2
```

```
2 1 2.6 3
2 2 6.1 16
2 3 5.8 6
;
proc glmpower data=Chow;
class Treat Guess;
model Outcome = Treat | Guess;
weight Weight;
contrast 'Inter-overall' Treat*Guess 1 0 -1 -1 0 1, Treat*Guess 0 1 -1 0 -1
1;
contrast 'Inter-tailored (ATE in 1(=T) vs. 2(=C))' Treat*Guess 1 -1 0 -1 1 0;
power
stddev = 4 5
ntotal = 148 98
power = .;
run;
```

References

Agresti A. Categorical Data Analysis. Wiley; 2013.

Arandjelović O. A new framework for interpreting the outcomes of imperfectly blinded controlled clinical trials. PLoS One. 2012; 7:e48984. [PubMed: 23236350]

Baethge C, Assall O, Baldessarini R. Systematic review of blinding assessment in randomized controlled trials in Schizophrenia and Affective Disorders 2000–2010. Psychotherapy and Psychosomatics. 2013; 82:152–160. [PubMed: 23548796]

Bang H. Random guess and wishful thinking are the best blinding scenarios. Contemporary Clinical Trials Communications. 2016; 3:117–121. [PubMed: 27822568]

Bang H, et al. Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol. Clinical Research and Regulatory Affairs. 2010; 27:42–51.

Bang H, Ni L, Davis CE. Assessment of blinding in clinical trials. Controlled Clinical Trials. 2004; 25:143–156. [PubMed: 15020033]

Bang H, Park J. Blinding in clinical trials: a practical approach. J Alternative and Complemtary Medicine. 2013; 19:367–369.

Boutron I, et al. Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. PLoS Med. 2007; 4:e61. [PubMed: 17311468]

Briggs A. Economic evaluation and clinical trials: size matters. The need for greater power in cost analyses poses an ethical dilemma. BMJ. 2000; 321:1362. [PubMed: 11099268]

Brinjikji W, et al. Investigational vertebroplasty efficacy and safetry trial: Detailed analysis of blinding efficacy. Radiology. 2010; 257:219–225. [PubMed: 20851942]

Brownell KD, Stunkard AJ. The double-blind in danger: untoward consequences of informed consent. Am J Psychiatry. 1982; 139:1487–1489. [PubMed: 6753613]

Castelloe JM, O'Brien RG. Power and sample size determination for linear models. Proceedings of the Twenty-sixth Annual SAS Users Group International Conference. 2001; 240

Chow SC, Shao J. Analysis of clinical data with breached blindness. Statistics in Medicine. 2004; 23:1185–1193. [PubMed: 15083477]

Cohen J. Things I have learned (so far). American Psychologist. 1990; 45:1304–1312.

COMMIT (ClOpidogrel and Metoprolol in Myocardial Infarction Trial) collaborative group. Addition of clopidogrel to aspirin in 45 852 patients with acute myocardial infarction: randomised placebo-controlled trial. Lancet. 2005; 366:1607–1621. [PubMed: 16271642]

CONSORT. 2010. http://www.consort-statement.org/consort-2010, last access on July 30, 2017

Crisp A. Blinding in pharmaceutical clinical trials: An overview of points to consider. Contemporary Clinical Trials. 2015; 43:155–163. [PubMed: 26044462]

Elston RC, Bush N. The hypotheses that can be tested when there are interactions in an analysis of variance model. Biometrics. 1964; 20:681–698.

Freed B, et al. Assessing blinding in trials of psychiatric disorders: A meta-analysis based on blinding index. Psychiatry Research. 2014; 219:241–247. [PubMed: 24930582]

Hertzberg V, et al. Use of dose modification schedules is effective for blinding trials of warfarin: evidence from the WASID study. Clinical Trials. 2008; 5:23–30. [PubMed: 18283076]

Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician. 2001; 55:1–6.

Hopton AK, Macpherson H. Assessing blinding in randomised controlled trials of acupuncture: challenges and recommendations. Chin J Integr Med. 2011; 17:173–176. [PubMed: 21359917]

Houweling A, et al. Blinding strategies in the conduct and reporting of a randomized placebo-controlled device trial. Clinical Trials. 2014; 11:547–552. [PubMed: 24902921]

Jadad A, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinical Trials. 1996; 17:1–12. [PubMed: 8721797]

James KE, et al. An index for assessing blindness in a multi-centre clinical trial: disulfiram for alcohol cessation - a VA cooperative study. Statistics in Medicine. 1996; 15:1421–1434. [PubMed: 8841652]

Jeong H, et al. The effect of rigorous study design in the research of autologous bone marrow-derived mononuclear cell transfer in patients with acute myocardial infarction. Stem Cell Research & Therapy. 2013; 4

Kolahi J, Bang H, Park J. Towards a proposal for assessment of blinding success in clinical trials: up-to-date review. Community Dentistry and Oral Epidemiology. 2009; 37:477–484. [PubMed: 19758415]

Mathieu E, et al. A theoretical analysis showed that blinding cannot eliminate potential for bias associated with beliefs about allocation in randomized clinical trials. Journal of Clinical Epidemiology. 2014; 67:667–671. [PubMed: 24767518]

Moroz A, et al. Blinding measured: a systematic review of randomized controlled trials of acupuncture. Evidence-Based Complementary and Alternative Medicine. 2013:708251. [PubMed: 23533515]

Muller KE, Peterson BL. Practical methods for computing power in testing the multivariate general linear hypothesis. Computational Statistics and Data Analysis. 1984; 2:143–158.

O'Brien RG. Using the SAS System to perform power analysis for log-linear models. The Proceedings of the Eleventh Annaul SAS Users Group International Conference. 1986

O'Brien RG, Shieh G. Pragmatic, unifying algorithm gives power probabilities for common F tests of the multivariate general linear hypothesis. The American Statistical Association Meetings. 1992

Park J, Bang H, Canette I. Blinding in clinical trials, time to do it better. Complementary Therapies in Medicine. 2008; 16:121–123. [PubMed: 18534323]

Park J, et al. Acupuncture for subacute stroke rehabilitation: a sham-controlled, subject- and assessor-blind, randomized trial. Archives of Internal Medicine. 2005; 165:2026–2031. [PubMed: 16186474]

Roy J. Randomized treatment-belief trials. Contemporary Clinical Trials. 2012; 33:172–177. [PubMed: 21989161]

Shin S, et al. Effectiveness and safety of electroacupuncture on poststroke urinary incontinence: study protocol of a pilot multicentered, randomized, parallel, sham-controlled trial. Evidence-Based Complementary and Alternative Medicine. 2016:5709295. [PubMed: 28042304]

Thompson SK. Sample size for estimating multinomial proportions. The American Statistician. 1987; 41:42–46.

Tortora RD. A note on sample size estimation for multinomial populations. The American Statistician. 1978; 32:100–102.

Vernon H. Chiropractic Manual Therapy and Neck Pain. 2017. https://clinicaltrials.gov/ct2/show/NCT01772966, Last accessed on July 30, 2017

Vernon H, et al. Retention of blinding at follow-up in a randomized clinical study using a sham-control cervical manipulation procedure for neck pain: secondary analyses from a randomized clinical study. J of Manipulative and Physiological Therapeutics. 2013; 36:522–526.

Walter S, Awasthi S, Jeyaseelan L. Pre-trial evaluation of the potential for unblinding in drug trials: a prototype example. Contemporary Clinical Trials. 2005; 26:459–468. [PubMed: 16054578]

Wilsey B, Deutsch R, Marcotte TD. Maintenance of blinding in clinical trials and the implications for studying analgesia using cannabinoids. Cannabis and Cannabinoid Research. 2016; 1:139–148. [PubMed: 28861490]

Wright S, Duncombe P, Altman DG. Assessment of blinding to treatment allocation in studies of a cannabis-based medicine (Sativex®) in people with multiple sclerosis: a new approach. Trials. 2012; 13:1–11. [PubMed: 22214287]

Zhang Z, et al. A causal model for joint evaluation of placebo and treatment-specific effects in clinical trials. Biometrics. 2013; 69:318–327. [PubMed: 23432119]

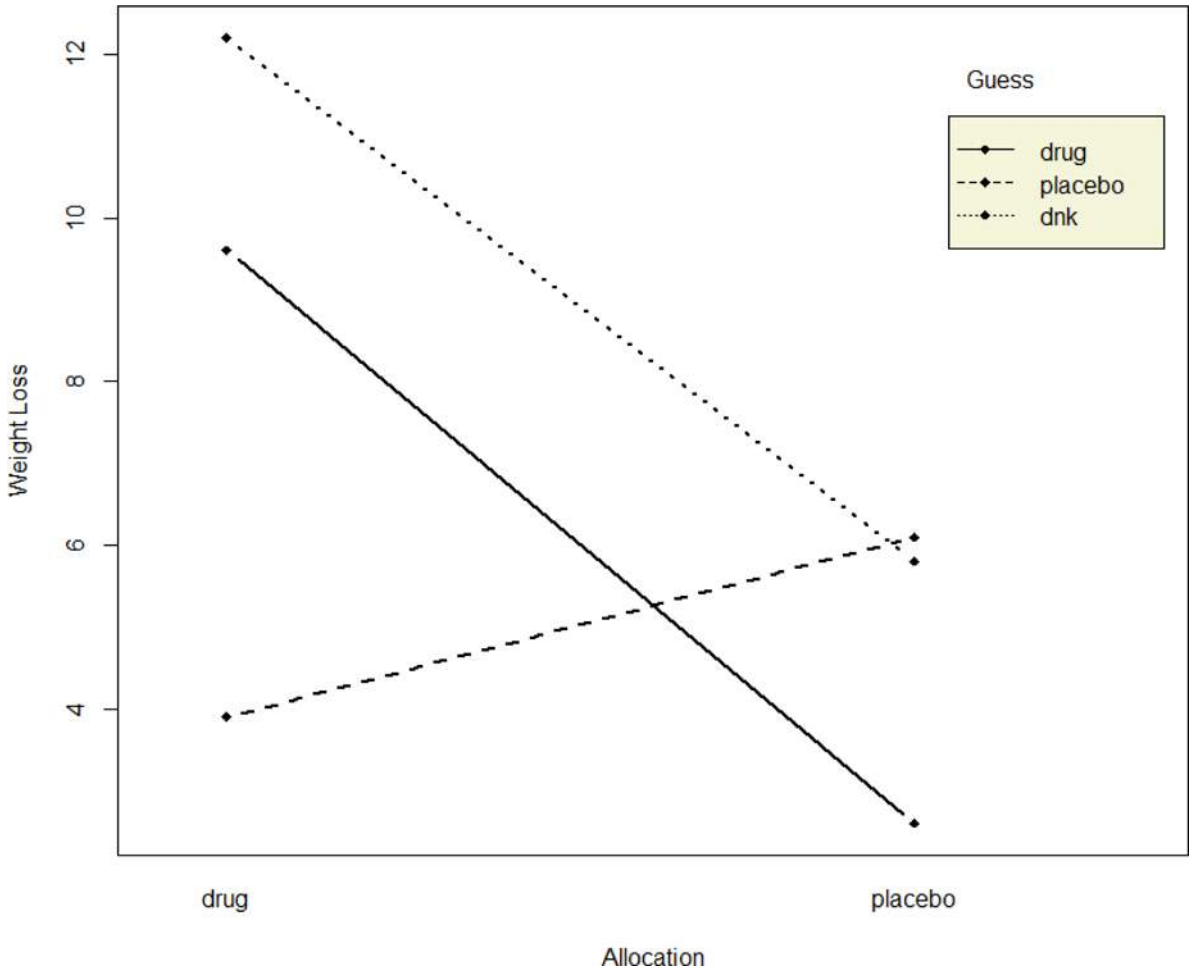

Figure 1. Clinical outcome by allocation and guess status

Created using data from Brownell and Stunkard (1982) and Chow and Shao (2004). drug: Guess T; placebo: Guess C; dnk: Guess 'Don't know'.

Table 1

Notations used to summarize typical blinding data

	Guess, count			Total
Allocation	**1 (=T)**	**2 (=C)**	**3 (=Don't know)**	**Sample size**
Treatment (T)	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$ $(= n_{T\cdot})$
Control (C)	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$ $(= n_{C\cdot})$

Table 2

Blinding data from Vernon et al. (2013) and Sample size/Power

a. Blinding data

	Guess			Total
Allocation	**Real**	**Sham**	**Don't know**	**Sample size**
Real	16	9	7	32
Sham	9	15	8	32

b. Sample size and power ($\alpha = 0.05$)

N	Power of Pearson Chi-square	Power of LR
20	0.14	0.14
50	0.30	0.30
100	0.56	0.55
176	0.80	0.81
200	0.85	0.86
300	0.96	0.96

Table 3

Sample size ($n_{T.}$) for the estimation of BI in treatment arm ($\alpha = 0.05$)

$BI_T = P_{1|T} - P_{2|T}$	d=0.2	d=0.1
$BI_T = 0$		
0.1-0.1	20	77
0.2-0.2	39	154
0.3-0.3	58	231
0.4-0.4	77	308
0.5-0.5 *	97	385
$BI_T = 0.1$		
0.1-0.0	9	35
0.2-0.1	28	115
$BI_T = 0.1, 0.2, 0.3, 0.4$		
0.3-0.2/0.3-0.1	48/35	189/139
0.4-0.3/0.4-0.2/0.4-0.1	67/54/40	266/216/158
0.5-0.4/0.5-0.3/0.5-0.2/0.5-0.1	86/73/59/43	342/292/235/170

$P_{1|T}$=expected proportion of persons who guessed 1 (=T).

$P_{2|T}$=expected proportion of persons who guessed 2 (=C).

BI_T denotes blinding index and n_T. denotes sample size in Treatment arm.

* The required sample size in this case is equivalent to a conservative sample size in a 2×2 format (without "Don't Know" category), $n=z^2_{\alpha/2}/d^2$.

Table 4

Blinding and outcome data from Brownell and Stunkard (1982) and Sample size/Power

a. Blinding data				
	Guess, count			**Total**
Allocation	**Active drug**	**Placebo**	**Don't know**	**Sample size**
Active drug	19	3	2	24
Placebo	3	16	6	25

b. Mean weight loss (kg) in subgroups defined by allocation and guess				
	Guess			
Assignment	**Active drug**	**Placebo**	**Don't know**	**Overall**
Active drug	9.6	3.9	12.2	9.1
Placebo	2.6	6.1	5.8	5.6

c. Sample size and power ($\alpha = 0.05$)			
Source	σ	**N**	**Power**
Main effect of guess (df=2)	4	148	0.860
	4	98	0.682
	5	148	0.668
	5	98	0.481
Overall interaction (df=2)	4	148	0.994
	4	98	0.948
	5	148	0.942
	5	98	0.806
Tailored interaction (df=1) ATE among guess=T vs. ATE among guess=C	4	148	0.994
	4	98	0.954
	5	148	0.947
	5	98	0.829

df: degrees of freedom; ATE: average treatment effect.