

A New Method for Choosing Sample Size for Confidence Interval–Based Inferences

Michael R. Jiroutek,^{1,*} Keith E. Muller,² Lawrence L. Kupper,²
and Paul W. Stewart²

¹Bristol-Myers Squibb Pharmaceutical Research Institute, 5 Research Parkway, Wallingford, Connecticut 06492-7660, U.S.A.

²Department of Biostatistics, University of North Carolina CB 7420 McGavran-Greenberg, Chapel Hill, North Carolina 27599-7420, U.S.A.

**email:* michael.jiroutek@bms.com

SUMMARY. Scientists often need to test hypotheses and construct corresponding confidence intervals. In designing a study to test a particular null hypothesis, traditional methods lead to a sample size large enough to provide sufficient statistical power. In contrast, traditional methods based on constructing a confidence interval lead to a sample size likely to control the width of the interval. With either approach, a sample size so large as to waste resources or introduce ethical concerns is undesirable. This work was motivated by the concern that existing sample size methods often make it difficult for scientists to achieve their actual goals. We focus on situations which involve a fixed, unknown scalar parameter representing the true state of nature. The width of the confidence interval is defined as the difference between the (random) upper and lower bounds. An event *width* is said to occur if the observed confidence interval width is less than a fixed constant chosen *a priori*. An event *validity* is said to occur if the parameter of interest is contained between the observed upper and lower confidence interval bounds. An event *rejection* is said to occur if the confidence interval excludes the null value of the parameter. In our opinion, scientists often implicitly seek to have all three occur: width, validity, and rejection. New results illustrate that neglecting rejection or width (and less so validity) often provides a sample size with a low probability of the simultaneous occurrence of all three events. We recommend considering all three events simultaneously when choosing a criterion for determining a sample size. We provide new theoretical results for any scalar (mean) parameter in a general linear model with Gaussian errors and fixed predictors. Convenient computational forms are included, as well as numerical examples to illustrate our methods.

KEY WORDS: Confidence interval; Power; Rejection; Sample size; Validity; Width.

1. Introduction

1.1 Motivation

Many statisticians and scientists strongly prefer confidence intervals over hypothesis tests. Much of the appeal arises from the ability of confidence intervals to help quantify the magnitude of an effect in units of scientific interest. Unfortunately, existing methods for choosing a sample size to compute a confidence interval often fail to address important scientific goals.

For example, Pisano et al. (2002) conducted a study to compare mammography displays. Traditionally, radiologists have read mammograms on film (hardcopy). Recently developed digital mammography equipment allows display on a computer screen (softcopy). In order to adopt the use of softcopy images, the time required to read a mammogram needs to be considered, in addition to image quality. In this study, radiologists were asked to read under both modalities in order to determine if the mean reading times differ substantially.

Such investigators often ask, “How many subjects are needed to have a high probability of producing a confidence

interval for the parameter of interest with width no greater than a fixed constant?” This question is usually easy to answer, given independent observations from distributions of assumed known structure (e.g., Gaussian). However, in many situations, the question is incomplete. In addition to desiring a narrow confidence interval for the true mean time difference, the scientists in our example were also very interested in knowing whether reading softcopy is faster or slower than reading hardcopy. That is, they were also interested in the rejection of the null hypothesis of no difference in true mean reading times.

Consider a fixed, unknown scalar parameter, θ , representing the true state of nature, with corresponding null value $\theta_0 < \theta$. With L and U as the lower and upper (random) bounds, confidence interval width is defined as $U - L$. An event *width* is defined as $U - L \leq \delta$, for fixed $\delta > 0$ chosen *a priori*. An event *validity* is defined as $L \leq \theta \leq U$. An event *rejection*, of the null hypothesis that $\theta = \theta_0$, is said to occur if the observed interval excludes θ_0 . As phrased in the question

above, only the *width* of the interval is considered, while *rejection* and *validity* have been neglected.

The new methods differ from previous work by *simultaneously* considering width, validity, and rejection to choose a sample size. We will argue that the best question is often “Given validity, how many subjects are needed to have a high probability of producing a confidence interval that correctly does not contain the null value when the null hypothesis is false and has a width no greater than δ ?” Addressing this question will lead to sample sizes that are more likely to achieve desired scientific goals than those chosen with traditional methods.

1.2 Notation

All results are presented in terms of a scalar (expected value) parameter in the general linear multivariate model (GLMM), assuming fixed predictors. The general notation includes a wide range of special cases: one-sample *t*-test, two-sample *t*-test, paired-data *t*-test, and planned scalar contrasts in univariate, multivariate, or repeated measures analysis of variance (ANOVA). The notation is essentially from Muller et al. (1992) and is summarized in Table 1.

Lowercase bold always indicates a (column) vector, while uppercase bold indicates a matrix. Whenever random variable and matrix notation conflict, matrix notation will dominate. Detailed information about all random variables discussed in this article can be found in Kotz, Balakrishnan, and Johnson (2000), and Johnson, Kotz, and Balakrishnan (1994, 1995).

For fixed and known design matrix \mathbf{X} , fixed unknown parameter matrix \mathbf{B} , observed responses \mathbf{Y} , and unobserved errors \mathbf{E} , the assumed model is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{1}$$

with rows of \mathbf{E} independent and $\text{row}_i(\mathbf{E})' \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, which indicates that $\text{row}_i(\mathbf{E})'$ is length p and follows a normal distribution with mean $\mathbf{0}$ and covariance matrix Σ . The usual estimators are $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\hat{\Sigma} = \mathbf{Y}'\{\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{Y}/\nu_e$, where $\nu_e = N - r$ is the error degrees of freedom (d.f.) and $r = \text{rank}(\mathbf{X})$. The associated general linear hypothesis (GLH) about $\Theta = \mathbf{C}\mathbf{B}\mathbf{U}$ can be stated

$$H_0 : \mathbf{C}\mathbf{B}\mathbf{U} = \Theta_0, \tag{2}$$

Table 1
Definitions of matrices

Symbol	Size	Definition and properties
\mathbf{X}	$N \times q$	Fixed, known design matrix
\mathbf{B}	$q \times p$	Primary parameters (means)
\mathbf{C}	$a \times q$	Between-subject contrasts
\mathbf{U}	$p \times b$	Within-subject contrasts
$\Theta = \mathbf{C}\mathbf{B}\mathbf{U}$	$a \times b$	Secondary parameters
Θ_0	$a \times b$	Parameter null values
Σ	$p \times p$	Covariance matrix of $\text{row}_i(\mathbf{E})'$
$\Sigma_* = \mathbf{U}'\Sigma\mathbf{U}$	$b \times b$	Covariance matrix of $\text{row}_i(\mathbf{E}\mathbf{U})'$
$\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$	$a \times a$	Middle matrix
$\Delta = (\Theta - \Theta_0)' \times \mathbf{M}^{-1}(\Theta - \Theta_0)$	$b \times b$	Unscaled ncentrality

for fixed and known Θ_0 ($a \times b$). Only *testable* hypotheses are considered, which require full rank Σ_* , \mathbf{M} , and \mathbf{U} and $\mathbf{C} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})$.

The special case $a = b = 1$ implies that the $a \times b$ secondary parameter Θ , the $a \times b$ known constant Θ_0 , and the $b \times b$ covariance matrix Σ_* become the scalars θ , θ_0 , and σ_*^2 , respectively. In turn, all univariate and multivariate repeated measures tests provide the same p-value. Define $\hat{\Delta} = (\hat{\Theta} - \Theta_0)' \times \mathbf{M}^{-1}(\hat{\Theta} - \Theta_0)$. The statistic to test the hypothesis in (2) may be computed as

$$\begin{aligned} F &= \frac{\text{trace}(\hat{\Delta})/(ab)}{\text{trace}(\hat{\Sigma}_*)/b} \\ &= \frac{(\hat{\theta} - \theta_0)'m^{-1}(\hat{\theta} - \theta_0)/1}{\hat{\sigma}_*^2/1} \\ &= \frac{(\hat{\theta} - \theta_0)^2}{\hat{\sigma}_*^2 m}, \end{aligned} \tag{3}$$

where m is the scalar version of the middle matrix \mathbf{M} defined in Table 1 and the simplifications arise from the restriction $a = b = 1$.

1.3 Example Details

The process of planning a follow-up study to Pisano et al. (2002) will illustrate the new methods. The randomness of the mean and variance estimators makes any sample size choice based on such estimators random. The desire to account for this randomness leads naturally to the desire to create a confidence interval around the (estimated) sample size. Taylor and Muller (1995, 1996), and Muller and Pasour (1997), derived exact methods for creating such confidence intervals in the context of power analysis for any general linear univariate model with fixed predictors. Including a careful and complete treatment of methods needed to account for such random values would considerably lengthen the present article. Hence, for the sake of brevity, we reserve that discussion for a future article.

In planning the Pisano et al. (2002) study, the scientists felt that radiologists would tolerate the disadvantage of an increase in true mean reading time of up to 25% in order to gain the many advantages of softcopy over hardcopy display. Experience with similar studies led us to expect that log response time would approximately follow a Gaussian distribution. Hence, the model was formulated in terms of the mean difference of the logarithms of reading times. With t_{hi} and t_{si} the random observed viewing times for reader i for hard and softcopy, respectively, it follows that $y_i = \log_{10}(t_{hi}/t_{si}) = \log_{10}(t_{hi}) - \log_{10}(t_{si})$. The model simplifies to $\mathbf{y} = \mathbf{1}_N\beta + \mathbf{e}$, with $\beta = \mathcal{E}(y_i)$. The hypothesis of interest has $\theta_0 = 0$ and $\theta = 1 \cdot \beta \cdot 1$, which reduces (2) to $H_0 : \theta = 0$. No information about the variance in reading time differences was available before the study began. A sample size of eight radiologists was chosen (with each radiologist reading a softcopy and hardcopy mammogram) in order to control costs while still hopefully providing a defensible variance estimate. It was expected that a subsequent study would be conducted, if necessary, to achieve more precise conclusions.

The paired-data analysis led to an estimated mean difference (hardcopy minus softcopy) of 0.076 \log_{10} seconds of

viewing time, with estimated error variance 0.012. Properties of the logarithmic transformation allow noting that the observed difference corresponds to an approximately 16% *reduction* in median reading time (softcopy better than hardcopy). In order to examine the sensitivity of sample size to the choice of inputs, we considered 10, 20, and 40% differences in true mean viewing time. Corresponding \log_{10} scale widths of 0.046, 0.097, and 0.222 lead to recruiting 106, 29, or 9 radiologists, respectively, based on confidence interval width alone.

The null hypothesis of interest is that there is a true mean difference of zero between hard and softcopy (\log_{10}) reading times. Since softcopy images may take more or less time, a two-sided test is required. Based on power considerations alone, assuming a true mean difference of 0.076 \log_{10} seconds and a target power of at least 0.9 for a paired-data t -test leads to using 24 radiologists.

The calculations dramatically illustrate the risks of what Muller et al. (1992) described (in the context of power analysis) as a misalignment of sample size rule and scientific objective. In the current example, sample size could be either more than four times too large (106 vs. 24) or roughly three times too small (9 vs. 24) when using a width criterion rather than a power criterion. The choice of criterion depends entirely on the scientific objective.

In our experience, scientists usually fail to control both power and width criteria, despite computing a confidence interval *and* conducting a test of the null hypothesis at the end of the study. We propose to resolve the conflict among sample size rules by requiring the sample size to meet both power and width criteria conditional on a validity criterion, resulting in the alignment of sample size rule and scientific objective. The impact of seeking a high probability of achieving a valid confidence interval of width no more than δ , while also requiring a high rejection probability, is the focus of this article.

1.4 Literature Review

All current sample size methods for confidence intervals are based on some combination of two objectives: validity and width. Following the Neyman-Pearson tradition, define θ as the fixed, unknown parameter of interest representing the true state of nature, θ_0 as the null (comparison) value, U as the upper (random) interval bound, L as the lower (random) interval bound, and assume $\theta > \theta_0$. The event validity (V) occurs if the observed interval contains the parameter of interest, namely, $\hat{L} \leq \theta \leq \hat{U}$, so that

$$\Pr\{V\} = \Pr\{L \leq \theta \leq U\}. \quad (4)$$

Setting $\Pr\{V\} = 1 - \alpha$, with α fixed *a priori*, $[L, U]$ is said to provide an exact $(1 - \alpha)$ -size confidence interval for θ . The term “validity” is in some sense misleading. We consider only valid procedures for computing confidence intervals, in the sense that all have a confidence coefficient of 95%. The (random) confidence interval is inherently valid regardless of whether or not its realization happens to capture the true value of the parameter. However, we use the term “validity” to describe whether or not the realization of the random confidence interval happens to capture the true value of the parameter.

Although some basic assumptions differ from those in the Neyman-Pearson tradition, current Bayesian methodol-

ogy also targets validity (conditional on the observed data) as the objective function for confidence intervals. However, Bayesians are allowed a more intuitive interpretation of a confidence interval, namely, the probability that the population parameter is between the observed realizations of the (random) L and U , given the observed data, is at least $1 - \alpha$. See Carlin and Louis (2000, Section 2.3.2, p. 35) for a fully Bayesian treatment of confidence (probability) intervals.

With $\delta > 0$ constant and fixed *a priori*, the event width (W) occurs if $\hat{U} - \hat{L} \leq \delta$, so that

$$\Pr\{W\} = \Pr\{U - L \leq \delta\}. \quad (5)$$

Kupper and Hafner (1989) noted that some popular sample size formulas for confidence intervals, which seek to control width, may poorly approximate the sample size needed due to the use of large sample approximations in lieu of exact small-sample results.

Lehmann (1959) stated, “there is no merit in short intervals that are far away from the true θ ,” suggesting that there is little reason to control the width of a confidence interval which does not have $\Pr\{V\} \geq 1 - \alpha$. Formalizing this idea, Beal (1989) advocated determining sample sizes using the conditional probability

$$\Pr\{W | V\} = \Pr\{U - L \leq \delta | L \leq \theta \leq U\} = \frac{\Pr\{W \cap V\}}{\Pr\{V\}}. \quad (6)$$

Beal concluded that realizations of confidence intervals which happen to include the true parameter tend to be slightly wider than confidence interval realizations in general (unconditionally). At about the same time, Hsu (1989) independently discussed $\Pr\{W \cap V\}$ in the multiple-comparison setting, presenting the two-treatment situation as a special case. Wang and Kupper (1997) and Pan and Kupper (1999) extended the width and the width given validity criteria to two-population and multiple-comparison settings for Gaussian data, while treating confidence interval width as random.

Bristol (1989) compared sample sizes based on $\Pr\{W\}$ to those based on power. He found comparisons difficult, since $\Pr\{W\}$ is not directly related to power. He had no clear preference for either method, except to note that the method used should align with the analysis goal.

While not the main focus of the work presented here, power analysis does play an important role. See Muller et al. (1992) for a review of power analysis in the GLMM.

Equivalence and noninferiority tests are special cases of hypothesis testing. Various connections between methods for confidence intervals, and methods for equivalence and noninferiority studies have been investigated. See Hsu et al. (1994), Bauer and Kieser (1996), Chow and Liu (2000), and Rashid (2000) for further information.

Cesana, Reina, and Marubini (2001) recommended controlling both power and confidence interval expected width when choosing a sample size for comparing a binomial proportion to a reference value. Their goals agree very closely with ours. In contrast to their one-sample binomial results, the new results here apply to any scalar hypothesis in a GLMM, and add the requirement that the confidence interval contains the parameter with a high probability.

2. New Results

2.1 Logic behind the Approach

The new results are founded on the premise that addressing both power and confidence interval criteria simultaneously will lead to the best choice of sample size for statistical inferences based on confidence intervals. All existing methods for the GLMM address rejection alone, width alone, or width given validity. Solving the problem in the GLMM framework allows developing a single approach that applies to a wide variety of common designs.

The new methods derived in this article were motivated by the following premise. Rules for choosing sample size for studies using confidence intervals for statistical inference have traditionally focused on controlling width alone. However, as Lehmann (1959), and then Beal (1989) and Hsu (1989), argued, confidence interval width should be controlled conditional on validity. Since confidence intervals ideally *exclude* the null value when the the alternative is true, which implies rejecting the corresponding null hypothesis, rejection should be considered simultaneously with width and validity.

2.2 Concept of Rejection

We define rejection, denoted R , as the event that the confidence interval does not contain the null value. Rejection is a third property which can be used to choose sample sizes for confidence interval–based inferences. Having computed a confidence interval, a data analyst may conduct a hypothesis test by observing whether or not the interval excludes the null value. For a two-sided test of $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$, the probability of the event rejection can be written as

$$\Pr\{R\} = \Pr\{(U < \theta_0) \cup (\theta_0 < L)\}. \quad (7)$$

In the special case of a one-sided hypothesis test of $H_0 : \theta = \theta_0$ vs. $H_a : \theta > \theta_0$ ($\theta < \theta_0$), $\Pr\{R\}$ reduces to $\Pr\{\theta_0 < L\}$ ($\Pr\{U < \theta_0\}$). The (unconditional) definition of power (the probability of rejecting the null hypothesis) and $\Pr\{R\}$ then coincide exactly. See Leventhal and Huynh (1996) for a related discussion.

2.3 An Exact Expression for $\Pr\{(W \cap R) | V\}$

For a two-sided situation, sample size may be chosen to control

$$\begin{aligned} \Pr\{(W \cap R) | V\} \\ = \Pr\{[(U - L \leq \delta) \cap (U < \theta_0 \cup \theta_0 < L)] | (L \leq \theta \leq U)\}. \end{aligned} \quad (8)$$

In words, $\Pr\{(W \cap R) | V\}$ is the probability that the width of an interval is less than a fixed constant and the null hypothesis is rejected, given that the interval contains the true parameter.

Varying the form of hypothesis test and confidence interval desired leads to several special cases of $\Pr\{(W \cap R) | V\}$. In practice, a two-sided hypothesis test and a two-sided confidence interval (2s test/2s CI) would typically be used together, although a one-sided hypothesis test might be paired with a one- or two-sided confidence interval (1s test/1s CI; 1s test/2s CI). The following theorem and corollaries provide expressions for $\Pr\{(W \cap R) | V\}$ and related probabilities for the GLMM framework. See the Appendix for all proofs.

THEOREM: With σ_*^2 , ν_e and m as defined in Section 1.3, let $f_{\text{crit}} = F_F^{-1}(1 - \alpha; 1, \nu_e)$ indicate the $(1 - \alpha)$ quantile of the cumulative distribution function (CDF) of a central F random variable with 1 as the numerator and ν_e as the denominator d.f. Also assume that $\theta > \theta_0$, $\delta (> 0)$ is the confidence interval width desired, $\theta_d = \theta - \theta_0$, $x_1 = \nu_e \delta^2 / (4\sigma_*^2 f_{\text{crit}} m)$, $c_1 = (f_{\text{crit}} / \nu_e)^{1/2}$, and $c_2 = \theta_d / (\sigma_*^2 m)^{1/2}$. Let $\Phi(\cdot)$ indicate the CDF of a standard normal variate and $f_{\chi^2}(x; \nu_e)$ the central chi-squared density function with ν_e d.f. For a 2s test/2s CI, with $a = b = 1$ (which insures a scalar parameter),

$$\begin{aligned} \Pr\{(W \cap R) | V\} \\ = \int_0^{x_1} [\Phi(c_1 x^{1/2}) - \Phi\{\max(c_1 x^{1/2} - c_2, -c_1 x^{1/2})\}] \\ \times \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx. \end{aligned} \quad (9)$$

COROLLARY 1: Assume $\theta > \theta_0$. For the one-sided test $H_0 : \theta = \theta_0$ vs. $H_a : \theta > \theta_0$ with $a = b = 1$, and a two-sided confidence interval, (9) still holds.

COROLLARY 2: Assume $\theta > \theta_0$. For the one-sided test $H_0 : \theta = \theta_0$ vs. $H_a : \theta > \theta_0$ with $a = b = 1$, and a lower one-sided confidence interval of the form $[L, \infty)$, the probability is

$$\begin{aligned} \Pr\{(W \cap R) | V\} \\ = \int_0^{x_1} \{\Phi(c_1 x^{1/2}) - \Phi(c_1 x^{1/2} - c_2)\} \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx. \end{aligned} \quad (10)$$

COROLLARY 3: Assume $\theta < \theta_0$. For the one-sided test $H_0 : \theta = \theta_0$ vs. $H_a : \theta < \theta_0$ with $a = b = 1$, and a two-sided confidence interval, the probability is

$$\begin{aligned} \Pr\{(W \cap R) | V\} \\ = \int_0^{x_1} [\Phi\{\min(-c_1 x^{1/2} - c_2, c_1 x^{1/2})\} - \Phi(-c_1 x^{1/2})] \\ \times \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx. \end{aligned} \quad (11)$$

COROLLARY 4: Assume $\theta < \theta_0$. For the one-sided test $H_0 : \theta = \theta_0$ vs. $H_a : \theta < \theta_0$ with $a = b = 1$ and considering an upper one-sided confidence interval of the form $(-\infty, U]$, the probability is

$$\begin{aligned} \Pr\{(W \cap R) | V\} \\ = \int_0^{x_1} \{\Phi(-c_1 x^{1/2} - c_2) - \Phi(-c_1 x^{1/2})\} \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx. \end{aligned} \quad (12)$$

COROLLARY 5: Alternate forms for Beal’s (1989) $\Pr\{W | V\}$, and Hsu’s (1989) $\Pr\{W \cap V\}$, can be immediately derived as special cases by eliminating rejection (R) for each of the one- and two-sided cases described above.

Three distinct equalities deserve mention: i) The symmetry of the normal distribution leads to the equivalence of the form of (9) in Corollary 1 and (11) in Corollary 3, which both involve a 1s test/2s CI; ii) A similar equivalence holds between (10) and (12), which both involve a 1s test/1s CI; iii) Requiring validity in $\Pr\{(W \cap R) | V\}$ disallows the “opposite” tail,

meaning that (9) holds for the situation described in Corollary 1. Some practical implications of these equivalencies are described in Section 5.

2.4 A Better Computational Form for $\Pr\{(W \cap R) | V\}$ in Equation (9)

In some cases, equation (9) leads to computational difficulties which can be avoided as follows. If $x_0 = \theta_d^2 x_1 / \delta^2$, then $c_1 x^{1/2} - c_2 = -c_1 x^{1/2}$. The strictly increasing function $c_1 x^{1/2} - c_2$ and strictly decreasing function $-c_1 x^{1/2}$ intersect at $x_0 = \theta_d^2 x_1 / \delta^2$. When $\theta > \theta_0$, c_1 and c_2 are both nonnegative; so when $x_1 > x_0$, $\max(c_1 x^{1/2} - c_2, -c_1 x^{1/2}) = c_1 x^{1/2} - c_2$; when $x_1 \leq x_0$, $\max(c_1 x^{1/2} - c_2, -c_1 x^{1/2}) = -c_1 x^{1/2}$. Thus,

$$\Pr\{(W \cap R) | V\} = \begin{cases} \int_0^{x_0} \{\Phi(c_1 x^{1/2}) - \Phi(-c_1 x^{1/2})\} \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx \\ + \int_{x_0}^{x_1} \{\Phi(c_1 x^{1/2}) - \Phi(c_1 x^{1/2} - c_2)\} \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx, & \delta > \theta_d; \\ \int_0^{x_1} \{\Phi(c_1 x^{1/2}) - \Phi(-c_1 x^{1/2})\} \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx, & \delta \leq \theta_d. \end{cases} \tag{13}$$

In the following, $d(x) = 0$ if $x \leq x_0$, while $d(x) = 1$ if $x > x_0$. The first two integrals in (13) can be combined and rewritten into a more computationally efficient form, yielding

$$\Pr\{(W \cap R) | V\} = \begin{cases} \int_0^{x_1} [\{1 - d(x)\} \{\Phi(c_1 x^{1/2}) - \Phi(-c_1 x^{1/2})\} \\ + d(x) \{\Phi(c_1 x^{1/2}) - \Phi(c_1 x^{1/2} - c_2)\}] \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx, & \delta > \theta_d; \\ \int_0^{x_1} \{\Phi(c_1 x^{1/2}) - \Phi(-c_1 x^{1/2})\} \frac{f_{\chi^2}(x; \nu_e)}{1 - \alpha} dx, & \delta \leq \theta_d. \end{cases} \tag{14}$$

Computing each case of $\Pr\{(W \cap R) | V\}$ and $\Pr\{W | V\}$ requires specifying the values for $\{\theta_d, \delta, \sigma_*^2, \nu_e, \alpha\}$. A scale-free (canonical) form for these parameters is $\{\theta_d/\sigma_*, \delta/\sigma_*, \nu_e, \alpha\}$ since, for $c > 0$, the sets $\{\theta_d, \delta, \sigma_*^2, \nu_e, \alpha\}$ and $\{c\theta_d, c\delta, c\sigma_*^2, \nu_e, \alpha\}$ yield identical results.

3. Numerical Results

3.1 Computational Methods

All programs were written in SAS/IML (SAS Institute, 1999). Exact numerical integration used the QUAD function. A limited set of simulations helped check the programming accuracy and also the original derivation. Direct numerical integration allowed computing over one hundred $\Pr\{(W \cap R) | V\}$ values per second on a 450 MHz PC.

Using equation (14) to compute values of $\Pr\{(W \cap R) | V\}$ near 1.0 and sample sizes greater than 300 led to numerical instability. Applying a quantile transform (Glueck and Muller, 2001) eliminated all numerical instability and added only a small percent increase in computation time. For this application, let $F_{\chi^2}(x; \nu_e)$ indicate a central chi-squared c.d.f. with ν_e d.f. and corresponding $(1 - \alpha)$ quantile

$F_{\chi^2}^{-1}(1 - \alpha; \nu_e)$. The actual transformation is $p = F_{\chi^2}(x; \nu_e)$, with $x = F_{\chi^2}^{-1}(p; \nu_e)$ and $dp = f_{\chi^2}(x; \nu_e) dx$. The bounds become 0 and $F_{\chi^2}(x_1; \nu_e)$.

3.2 Comparing Sample Sizes

Choosing to control $\Pr\{(W \cap R) | V\}$, $\Pr\{W | V\}$, $\Pr\{W\}$ or $\Pr\{R\}$ (i.e., power) as the design goal can dramatically affect the sample size required. The various results in this section will illustrate this important conclusion. In contrast, the “sidedness” (i.e., whether or not the test or confidence interval is one-sided or two-sided) has little effect on the resulting sample size. Therefore, detailed numerical results are reported only for the 2s test/2s CI, although all other cases were examined numerically (based on the corresponding theory in Section 2). In particular, the value of $\Pr\{W\}$ does not depend on the sidedness of the test and confidence interval, due to underlying mathematical relationships. This occurs because one-sided intervals control only half of the corresponding two-sided interval. Comparing the proofs in the Appendix for the one- and two-sided cases illustrates this point in detail. Furthermore, only slight numerical differences in $\Pr\{(W \cap R) | V\}$, $\Pr\{W | V\}$, and $\Pr\{R\}$ were noted for the conditions considered. Overall, the 1s test/2s CI and 1s test/1s CI probabilities differed from the 2s test/2s CI probabilities by no more than 0.025 for $\Pr\{(W \cap R) | V\}$, and by no more than 0.007 for $\Pr\{W | V\}$ and $\Pr\{R\}$.

Figure 1 contains nine plots of N , with \log_2 spacing and $\nu_e = N - r = N - 2$, versus the probability of achieving the desired event, for $\alpha = 0.05$, $\sigma_*^2 = 1$, $\delta \in \{0.5, 1.0, 1.5\}$, $\theta_d \in \{0.5, 1.0, 1.5\}$, and $\theta_0 = 0$ (note that $\theta_0 = 0$ implies $\theta = \theta_d$). In all computations and plots, $\Pr\{W | V\}$ and $\Pr\{W\}$ were virtually indistinguishable, never being more than 5% apart. Given the previously stated preference for $\Pr\{W | V\}$, $\Pr\{W\}$ was dropped from further consideration, and is not included in the plots.

The conditions in Figure 1 fall into two groups. Those on and above the diagonal (from upper left to lower right) have $\delta \leq \theta_d$, while those below the diagonal have $\delta > \theta_d$. If $\delta \leq \theta_d$, then $\Pr\{(W \cap R) | V\}$ and $\Pr\{W | V\}$ coincide in the plots and mathematically, as can be confirmed via proofs in the Appendix. Comparisons among plots clearly illustrate the dramatic impact that alignment or misalignment of target probability with scientific goals may have.

Table 2 provides additional detail for each plot in Figure 1. The sample sizes vary due to the choice of target probability,

Table 2
Sample size (N) for (i) $\Pr\{(W \cap R) | V\}$, (ii) $\Pr\{W | V\}$, and (iii) $\Pr\{R\}$; $\sigma_*^2 = 1$, $\theta_0 = 0$, $\alpha = 0.05$, and $\nu_e = N - r$, $r = 2$

δ	Prob.	$\theta_d = 0.5$			$\theta_d = 1.0$			$\theta_d = 1.5$		
		(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0.5	0.8	268	268	128	268	268	34	268	268	18
	0.9	276	276	172	276	276	46	276	276	22
1.0	0.8	124	74	128	74	74	34	74	74	18
	0.9	160	78	172	78	78	46	78	78	22
1.5	0.8	124	36	128	40	36	34	36	36	18
	0.9	160	40	172	44	40	46	40	40	22

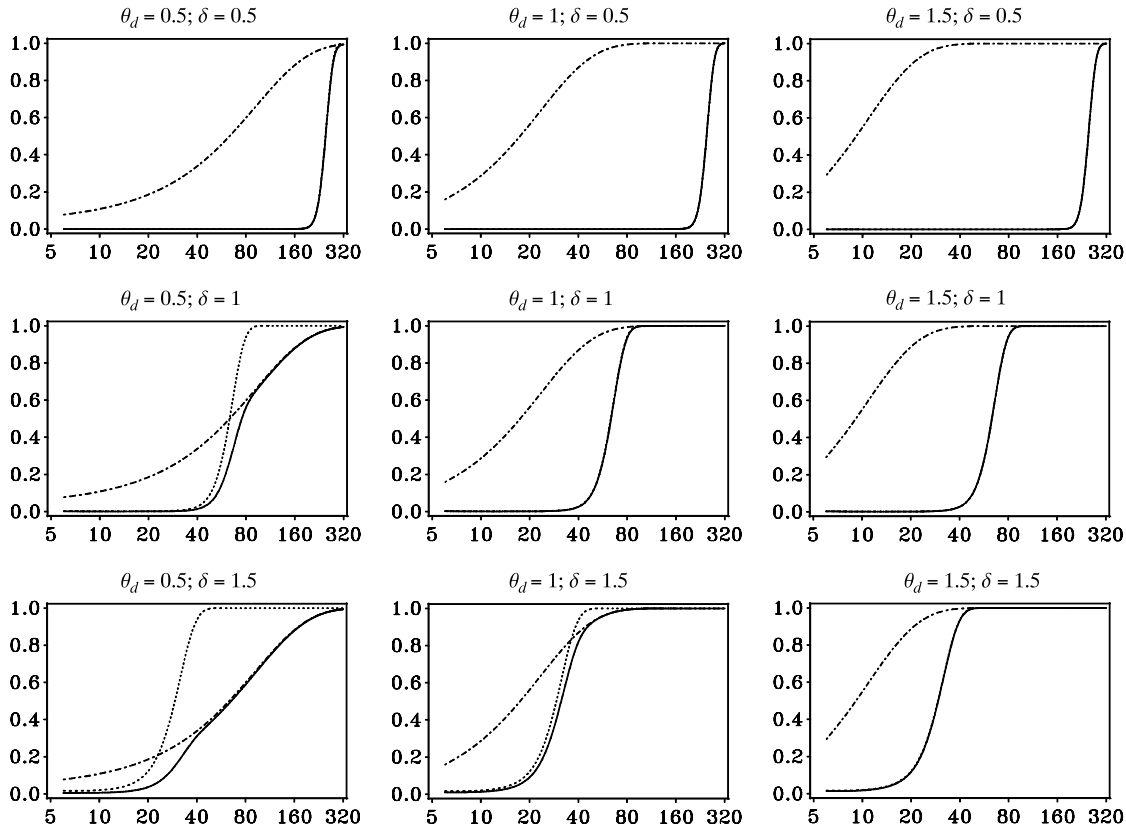


Figure 1. Event probabilities as a function of N with \log_2 spacing, $\nu_e = N - r$, $r = 2$, $\sigma_x^2 = 1$, $\theta_0 = 0$, and $\alpha = 0.05$. $\Pr\{(W \cap R) | V\}$: solid line; $\Pr\{R\}$: dashed line; $\Pr\{W | V\}$: dotted line.

$\Pr\{(W \cap R) | V\}$, $\Pr\{W | V\}$, or $\Pr\{R\}$, and the numeric value specified, either 0.8 or 0.9.

The major conclusion to be drawn from Figure 1 and Table 2 is that failure to align the event probability used to choose a sample size with the primary study endpoints can result in serious sample size errors. First, consider $\theta_d = 1.5$ and $\delta = 0.5$. Achieving $\Pr\{R\} \geq 0.90$ requires $N = 22$. However, achieving $\Pr\{(W \cap R) | V\} = \Pr\{W | V\} \geq 0.90$ requires $N = 276$ subjects! Second, consider the situation with $\theta_d = 0.5$ and $\delta = 1.5$. Achieving $\Pr\{R\} \geq 0.90$ requires $N = 172$, and to obtain $\Pr\{(W \cap R) | V\} \geq 0.90$ requires $N = 160$. In contrast, to have $\Pr\{W | V\} \geq 0.90$ requires only $N = 40$ subjects!

The sudden rise in probability that can be seen in the $\Pr\{(W \cap R) | V\}$ and $\Pr\{W | V\}$ curves, especially in the first row of plots in Figure 1, can be explained in two ways. First, the choice of \log_2 scale sample size was made to most effectively plot the $\Pr\{R\}$, $\Pr\{W | V\}$, and $\Pr\{(W \cap R) | V\}$ curves simultaneously. Unfortunately, this results in the $\Pr\{(W \cap R) | V\}$ and $\Pr\{W | V\}$ curves appearing to rise sharply at an arbitrary point. The choice of a different log base would flatten the curves, but make them more difficult to display on the same set of axes. Secondly, the sensitivity of the $\Pr\{(W \cap R) | V\}$ and $\Pr\{W | V\}$ curves to the choice of δ is reflected in the steep slopes of these curves. The impact of δ on these curves can also be seen by noticing the steepness of the $\Pr\{W | V\}$ and $\Pr\{(W \cap R) | V\}$ curves relative to the $\Pr\{R\}$ curve.

One last feature of Figure 1 deserves mention, although it is difficult to see given the size of the individual plots in the figure. Consider the curve for $\Pr\{(W \cap R) | V\}$ in the plot in the lower left corner and the same curve in the plot immediately above it. Neither has exactly the classical “S” shape commonly seen in sample size function curves. The $\Pr\{(W \cap R) | V\}$ curve in each plot is smooth in the technical sense in that it has a continuous first derivative and is also strictly monotone. Nonmonotone variation in the second derivative corresponds to the “bumpy” shape. The bumpy sections of the curves reflect the discord between the events rejection and width as each tries to dominate the calculation. It is not coincidence that the bumps occur in abscissa ranges where the inflection points occur for the $\Pr\{R\}$ and $\Pr\{W | V\}$ curves.

A number of features of Table 2 merit comment. Since $\Pr\{R\}$ is independent of δ , the sample size required to achieve the $\Pr\{R\}$ criterion is the same for any δ . Consider, for example, $\theta_d = 0.5$ and $\Pr\{R\} = 0.8$. The same sample size of 128 is required for $\delta = 0.5$, $\delta = 1.0$, and $\delta = 1.5$. Similarly, since $\Pr\{W | V\}$ is independent of θ_d , the sample sizes required using the $\Pr\{W | V\}$ criterion are constant for a fixed target probability (0.80 or 0.90), as θ_d changes across columns. As expected, the sample sizes for $\Pr\{(W \cap R) | V\}$ and $\Pr\{W | V\}$ are identical when $\theta_d \geq \delta$. Also, as δ increases, $\Pr\{(W \cap R) | V\}$ and $\Pr\{R\}$ essentially coincide. This occurs because as δ increases, the width

component of $\Pr\{(W \cap R) | V\}$ becomes less restrictive, increasing the relative role of rejection in the calculation. If $\delta = \infty$, then $\Pr\{(W \cap R) | V\} = \Pr\{R | V\}$, which is close to $\Pr\{R\}$ (for typical values of $\Pr\{V\}$, such as 0.95, which are near 1.0). Lastly, when $\delta = 1.5$, $\theta_d = 1.0$, and $\text{Prob.} = 0.8$, in Table 2, the sample size for $\Pr\{(W \cap R) | V\}$ is greater than that for both $\Pr\{W | V\}$ and $\Pr\{R\}$. This counterintuitive result reflects the impact of conditioning on the event validity (V). Recall that $\Pr\{(W \cap R) | V\} = \Pr\{W \cap R \cap V\} / \Pr\{V\}$ and note that $\Pr\{W \cap R \cap V\} \leq \Pr\{R\}$. For this particular situation, since the sample sizes for $\Pr\{(W \cap R) | V\}$, $\Pr\{W | V\}$, and $\Pr\{R\}$ are so close, the denominator ($\Pr\{V\} = 0.95$ for all cases in the figures provided) causes this seemingly paradoxical result.

3.3 How Should δ and θ_d be Chosen?

Although Figure 1 contains a great deal of information, it also raises a number of interesting questions. The interaction between δ and θ_d in the computation of $\Pr\{(W \cap R) | V\}$ yields sample sizes that are sensitive to the choice of each parameter, particularly δ . Figures 2 and 3, which are analogs to Figure 1, display event probabilities as a function of δ and θ_d , respectively. The figures were created to provide further guidance in the choice of δ and θ_d and give a more complete picture of the interaction between δ , θ_d , and N .

Figure 2 contains nine plots of δ , with \log_2 spacing, versus the probability of achieving the desired event, while Figure 3

contains nine plots of θ_d , with \log_2 spacing, versus the probability of achieving the desired event, both with $N \in \{20, 50, 100\}$ and $\nu_e = N - r = N - 2$. All other values remain the same as in Figure 1. Jointly examining the three figures allows one to form guidelines to handle the four dimensional problem, which requires specifying three of δ/σ_* , θ_d/σ_* , N and the probability of interest to determine the fourth. A range of N , or $\Pr\{(W \cap R) | V\}$, is typically specified, allowing $\Pr\{(W \cap R) | V\}$ or N to be computed across that range. The choice of δ and θ_d must be based on scientific, not statistical, principles.

The choice of δ is determined from scientific, monetary, temporal, and ethical considerations in much the same way that θ_d is for power analysis. A critical part of the consultation process with investigators is the elicitation of scientifically plausible values for δ and θ_d , and, in turn, their relative size. In particular, consider the Pisano et al. (2002) example. In that context, the choice of δ is determined largely by the practical consideration of the inconvenience to the radiologist. However, θ_d is controlled by the maximum tolerable (clinically useful) increase in reading time between hardcopy and softcopy.

We agree with Lenth's (2001) position that choice of sample size (e.g., analyses based on $\Pr\{(W \cap R) | V\}$, $\Pr\{W | V\}$, or $\Pr\{R\}$) should be cast in the units of the data, not in the abstract. Although Figures 1-3 serve as an excellent guide to determine sample size based on the new criterion,

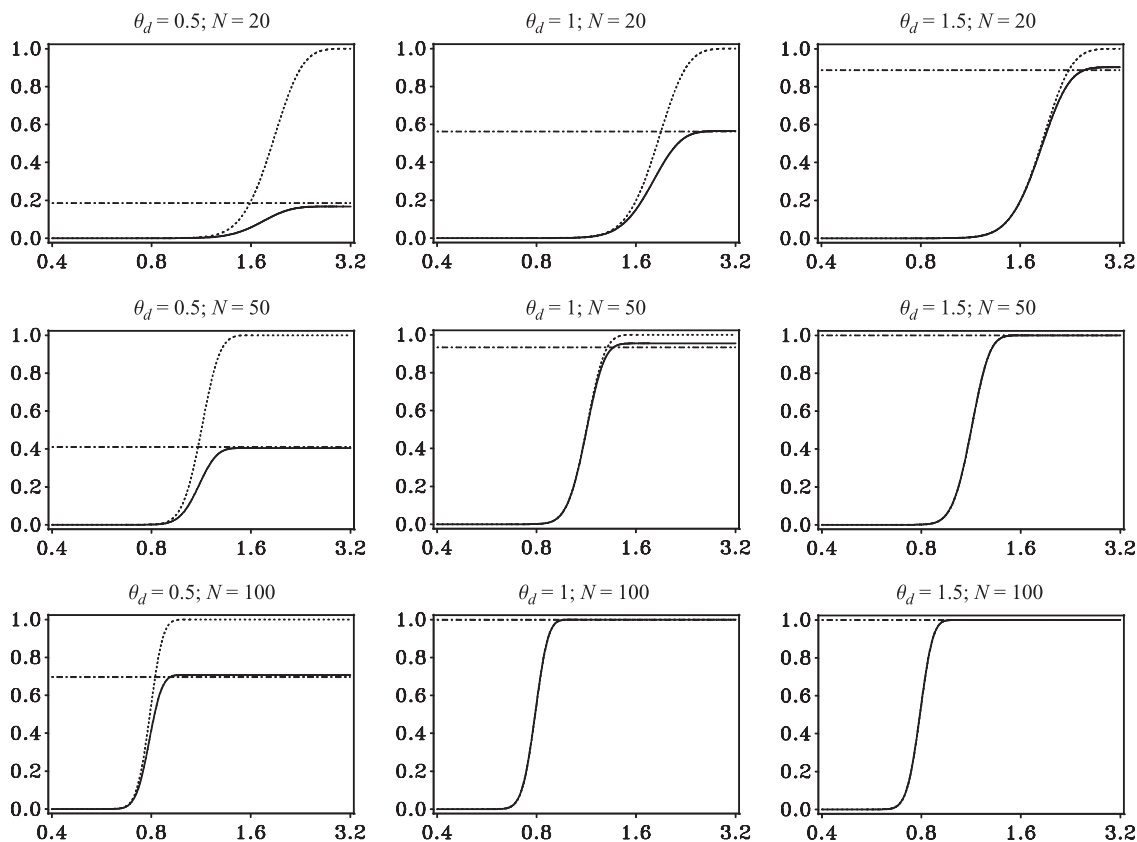


Figure 2. Event probabilities as a function of δ with \log_2 spacing, $\nu_e = N - r$, $r = 2$, $\sigma_*^2 = 1$, $\theta_0 = 0$, and $\alpha = 0.05$. $\Pr\{(W \cap R) | V\}$: solid line; $\Pr\{R\}$: dashed line; $\Pr\{W | V\}$: dotted line.

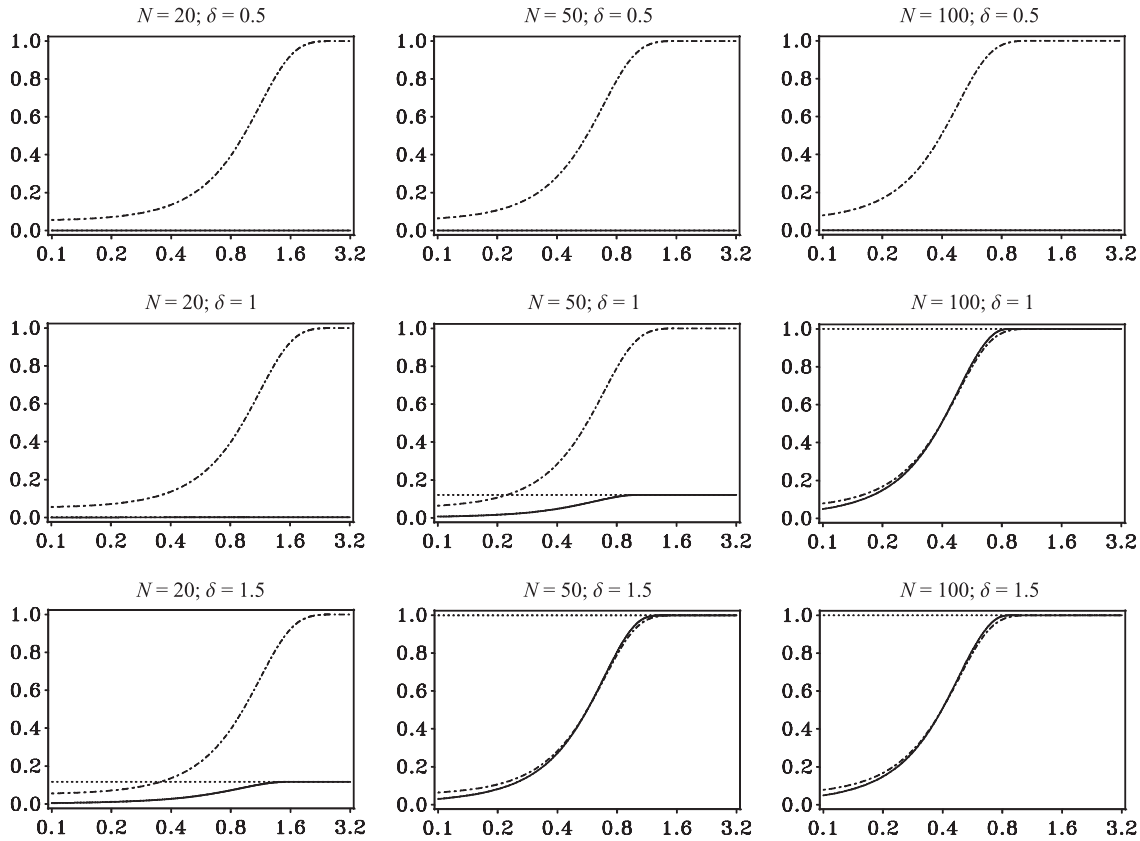


Figure 3. Event probabilities as a function of θ_d with \log_2 spacing, $\nu_e = N - r$, $r = 2$, $\sigma_*^2 = 1$, $\theta_0 = 0$, and $\alpha = 0.05$. $\Pr\{(W \cap R) | V\}$: solid line; $\Pr\{R\}$: dashed line; $\Pr\{W | V\}$: dotted line.

$\Pr\{(W \cap R) | V\}$, accurate analysis specific to a particular study should be completed with speculations for parameter values chosen for the study at hand.

There are several points worth mentioning about Figures 2 and 3. The relationship between $\Pr\{(W \cap R) | V\}$, $\Pr\{W | V\}$, and $\Pr\{R\}$ is complex, due to the interaction between the parameters θ_d , δ , and N . As described at the end of Section 2.2, $\Pr\{W \cap R \cap V\} \leq \Pr\{R\}$. Of course, similar reasoning implies $\Pr\{W \cap R \cap V\} \leq \Pr\{W \cap V\}$. However, conditioning on the event V explains why $\Pr\{(W \cap R) | V\}$ is not necessarily less than $\Pr\{R\}$, although the inequality $\Pr\{(W \cap R) | V\} \leq \Pr\{W | V\}$ always holds. These facts are illustrated by the figures.

It may require some thought to understand why horizontal lines occur in many of the plots in Figures 2 and 3. Since $\Pr\{R\}$ is independent of δ , $\Pr\{R\}$ is constant in each plot in Figure 2, but increases across each row as sample size increases. In Figure 3, $\Pr\{W | V\}$ is independent of θ_d and hence constant in each plot, but also increases across each row as sample size increases. Furthermore, in the top row of plots in Figure 3 and in the leftmost plot of the second row, note that $\Pr\{(W \cap R) | V\} = \Pr\{W | V\} = 0$. Since the inequality $\Pr\{(W \cap R) | V\} \leq \Pr\{W | V\}$ must always hold, and $\Pr\{W | V\}$ is constant in each plot in Figure 3, $\Pr\{W | V\} = 0$ implies $\Pr\{(W \cap R) | V\} = 0$. With $\Pr\{W | V\} > 0$ in the lower portion of Figure 3, $\Pr\{(W \cap R) | V\}$ approaches $\Pr\{W | V\}$ as θ_d increases. Since $\Pr\{(W \cap R) | V\}$ is jointly dependent on θ_d and δ , it is not constant in either figure.

4. Example Revisited

Consider again the Pisano et al. (2002) study introduced in Section 1.1. Table 3 contains sample sizes for \log_{10} -scale confidence interval widths corresponding to a reduction of 10, 20, or 40% in viewing time, with a desired target probability of at least 0.9. Table 3 further illustrates the potential for excessive or inadequate sample size due to misalignment.

Table 3
Softcopy study sample size (N) for $\theta_d = 0.076$, $\sigma_*^2 = 0.012$, $\theta_0 = 0$, $\alpha = 0.05$, and $\nu_e = N - r$, $r = 1$

δ	$\Pr\{R\}$	$\Pr\{W V\}$	$\Pr\{(W \cap R) V\}$
0.046	24	106	106
0.097	24	30	30
0.222	24	9	23

5. Discussion and Conclusions

Several conclusions arise from the four possible cases based on combining a one- or two-sided with a one- or two-sided confidence interval. The 1s test/1s CI and 2s test/2s CI combinations have obvious applications and have a simple relationship to each other. More precisely, for the Gaussian theory setting developed here, if α for the 1s test/1s CI method is half that for the 2s test/2s CI method, then the sample sizes chosen will be nearly the same. Combining a one-sided test with a two-sided interval seems both natural

and appealing. However, the hypothesis test size must be exactly twice the α for the confidence interval in order to avoid serious logical inconsistencies in the interpretation of the results for the symmetric distributions considered here. Finally, using a two-sided test with a one-sided interval has no logical appeal to us.

The examples in Figure 1 and Tables 2 and 3 illustrate the magnitude of error that can be made in study planning due to misaligning the sample size rule with the scientific goals. Furthermore, such errors can occur in either direction: choosing a sample size much smaller than necessary allows virtually no chance of achieving a successful outcome; alternately, choosing a sample size far larger than necessary may waste significant resources and create unnecessary risk to subjects. Scientists often seek to both test hypotheses and construct corresponding confidence intervals. Targeting $\Pr\{(W \cap R) | V\}$, rather than $\Pr\{R\}$, $\Pr\{W | V\}$, or $\Pr\{W\}$, helps achieve both goals in a single study, without undue cost or risk to subjects.

Defensible study design requires aligning the sample size rule with the scientific goal. The joint consideration of width, rejection and validity, especially in the calculation of $\Pr\{(W \cap R) | V\}$, is a new and practical tool for achieving such alignment. Either $\Pr\{W\}$ or $\Pr\{R\}$ may be emphasized by changing the relative sizes of δ and θ_d in $\Pr\{(W \cap R) | V\}$. In fact, $\Pr\{R | V\}$ and $\Pr\{W | V\}$ are special (limiting) cases of $\Pr\{(W \cap R) | V\}$.

ACKNOWLEDGEMENTS

Jiroutek's and Kupper's work is supported in part by NIEHS training grant 5-T32-ES07018. Muller's work is supported in part by NCI P01 CA47 982-04, NCI RO-1 CA095749-01A1, and NIAID 9P30 AI 50410. Stewart's work is supported in part by NICHD CFAR grant P30-HD-37260 and NIH GCRC grant 2 M01 RR00046-38.

RÉSUMÉ

Les scientifiques ont souvent besoin de tester des hypothèses et de construire les intervalles de confiance correspondant. En planifiant une étude pour tester une hypothèse nulle particulière, les méthodes traditionnelles conduisent à une taille d'échantillon assez grande pour fournir une puissance statistique suffisante. A l'opposé, les méthodes traditionnelles de construction d'intervalle de confiance conduisent à une taille d'échantillon appropriée pour contrôler la largeur de l'intervalle. Avec l'une ou l'autre des approches, une taille d'échantillon si grande qu'elle gaspille les ressources ou qu'elle introduise des questions éthiques n'est pas souhaitable. Ce travail a été motivé par le fait que les méthodes actuelles de recherche de taille d'échantillon rendent difficiles aux scientifiques l'atteinte de leurs objectifs. Nous nous centrons sur les situations qui impliquent un paramètre scalaire fixe mais inconnu représentant le vrai état de la nature. La largeur de l'intervalle de confiance est définie comme la différence entre les bornes (aléatoires) supérieures et inférieures. L'événement *largeur* est dit se réaliser si la largeur de l'intervalle de confiance observée est inférieure à une valeur constante fixée a priori. L'événement *validité* est dit se réaliser si le paramètre d'intérêt est situé entre les limites supérieure et inférieure observées de l'intervalle de confiance. L'événement *rejet* est dit se réaliser si l'intervalle de confiance exclut la valeur nulle du paramètre. Notre opinion est que les scientifiques recherchent souvent, de manière implicite, la

réalisation des ces trois événements: largeur, rejet et validité. De nouveaux résultats illustrent le fait de négliger le rejet ou la largeur (et à un moindre degré la validité) fournit souvent une taille d'échantillon avec une faible probabilité d'occurrence simultanée des trois événements. Nous recommandons de considérer ces trois événements simultanément pour déterminer une taille d'échantillon. Nous fournissons de nouveaux résultats théoriques pour n'importe quel paramètre scalaire (moyenne) dans un modèle linéaire général avec erreurs Gaussiennes et prédicteurs fixés. Des formes de calcul adaptées illustrent nos méthodes avec des exemples numériques.

REFERENCES

- Bauer, P. and Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83**(4), 934–937.
- Beal, S. L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* **45**, 969–977.
- Bristol, D. R. (1989). Sample sizes for constructing confidence intervals and testing hypotheses. *Statistics in Medicine* **8**, 803–811.
- Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edition. Tampa, Florida: Chapman and Hall/CRC.
- Cesana, B. M., Reina, G., and Marubini, E. (2001). Sample size for testing a proportion in clinical trials: A “two-step” procedure combining power and confidence interval expected width. *American Statistician* **55**(4), 288–292.
- Chow, S. and Liu, J. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd edition. New York: Marcel Dekker.
- Glueck, D. H. and Muller, K. E. (2001). On the expected values of sequences of functions. *Communications in Statistics—Theory and Methods* **30**, 363–369.
- Hsu, J. C. (1989). Sample size computation for designing multiple comparison experiments. *Computational Statistics and Data Analysis* **7**, 79–91.
- Hsu, J. C., Hwang, J. T. G., Liu, H. K., and Ruberg, S. J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika* **81**(1), 103–114.
- Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Volume 1, 2nd edition. New York: Wiley.
- Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Volume 2, 2nd edition. New York: Wiley.
- Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions*, Volume 1, 2nd edition. New York: Wiley.
- Kupper, L. L. and Hafner, K. B. (1989). How appropriate are popular sample size formulas? *American Statistician* **43**(2), 101–105.
- Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
- Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician* **55**, 187–193.
- Leventhal, L. and Huynh, C. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods* **1**(3), 278–292.

Muller, K. E. and Pasour, V. B. (1997). Bias in linear model power and sample size due to estimating variance. *Communications in Statistics—Theory and Methods* **26**(4), 839–851.

Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association* **87**(420), 1209–1226.

Pan, Z. and Kupper, L. L. (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine* **18**, 1475–1488.

Pisano, E. D., Cole, E. B., Kistner, E. O., et al. (2002). Interpretation of digital mammograms: A comparison of speed and accuracy of softcopy versus printed film display. *Radiology* **223**, 483–488.

Rashid, M. M. (2000). Rank-based procedures for non-inferiority and equivalence hypotheses in clinical trials when the centers are chosen at random. In *Proceedings of 2000 ASA Conference—Biopharmaceutical Section*, 127–132.

SAS Institute. (1999). *SAS/IML User’s Guide*, Version 8. Cary, North Carolina: SAS Institute.

Taylor, D. J. and Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *American Statistician* **49**(1), 43–47.

Taylor, D. J. and Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics—Theory and Methods* **25**, 1595–1610.

Wang, Y. and Kupper, L. L. (1997). Optimal samples sizes for estimating the difference in means between two normal populations treating confidence interval length as a random variable. *Communications in Statistics—Theory and Methods* **26**(3), 727–741.

Received June 2002. Revised February 2003.
Accepted March 2003.

APPENDIX

Proof of Theorem. Define a two-sided $100(1 - \alpha)\%$ confidence interval for θ by

$$\begin{aligned} 1 - \alpha &= \Pr \left\{ -f_{\text{crit}}^{1/2} \leq \frac{(\hat{\theta} - \theta)}{(\hat{\sigma}_*^2 m)^{1/2}} \leq f_{\text{crit}}^{1/2} \right\} \\ &= \Pr \{ \hat{\theta} - \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \leq \theta \leq \hat{\theta} + \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \} \\ &= \Pr \{ L \leq \theta \leq U \} \\ &= \Pr \{ |\theta - \hat{\theta}| \leq \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \} \\ &= \Pr \{ V \}, \end{aligned} \tag{A.1}$$

where $L = \hat{\theta} - \hat{\sigma}_*(f_{\text{crit}} m)^{1/2}$ and $U = \hat{\theta} + \hat{\sigma}_*(f_{\text{crit}} m)^{1/2}$. Additionally, with this notation,

$$\begin{aligned} \Pr \{ W \} &= \Pr \{ U - L \leq \delta \} \\ &= \Pr \{ 2\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \leq \delta \}, \end{aligned} \tag{A.2}$$

and

$$\begin{aligned} \Pr \{ R \} &= \Pr \{ (U < \theta_0) \cup (\theta_0 < L) \} \\ &= \Pr \{ [\hat{\theta} + \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} < \theta_0] \cup [\theta_0 < \hat{\theta} - \hat{\sigma}_*(f_{\text{crit}} m)^{1/2}] \} \\ &= \Pr \{ [\hat{\theta}_d < -\hat{\sigma}_*(f_{\text{crit}} m)^{1/2}] \cup [\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} < \hat{\theta}_d] \}, \end{aligned} \tag{A.3}$$

where $\hat{\theta}_d = \hat{\theta} - \theta_0$. Define $X = \nu_e(\hat{\sigma}_*^2/\sigma_*^2)$, $x_1 = \nu_e\delta^2/(4\sigma_*^2 f_{\text{crit}} m)$, $c_1 = (f_{\text{crit}}/\nu_e)^{1/2}$, $\theta_d = \theta - \theta_0 > 0$, $c_2 = \theta_d/(\sigma_*^2 m)^{1/2}$ and note that $\hat{\theta}_d \sim \mathcal{N}(\theta_d, \sigma_*^2 m)$ and $X \sim \chi^2(\nu_e)$, so that X follows a central chi-squared distribution with ν_e d.f. Since $Z = (\hat{\theta} - \theta)/(\sigma_*^2 m)^{1/2} \sim \mathcal{N}(0, 1)$, equation (A.1) can be rewritten as

$$\begin{aligned} |\theta - \hat{\theta}| &\leq \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \Leftrightarrow \\ \{ \theta - \hat{\theta} \leq \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \} &\cap \{ -(\theta - \hat{\theta}) \leq \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \} \Leftrightarrow \\ \left\{ \frac{\theta - \hat{\theta}}{(\sigma_*^2 m)^{1/2}} \leq \frac{\hat{\sigma}_*(f_{\text{crit}} m)^{1/2}}{(\sigma_*^2 m)^{1/2}} \right\} &\cap \left\{ \frac{\hat{\theta} - \theta}{(\sigma_*^2 m)^{1/2}} \leq \frac{\hat{\sigma}_*(f_{\text{crit}} m)^{1/2}}{(\sigma_*^2 m)^{1/2}} \right\} \Leftrightarrow \\ \left\{ -Z \leq \left(\frac{X f_{\text{crit}}}{\nu_e} \right)^{1/2} \right\} &\cap \left\{ Z \leq \left(\frac{X f_{\text{crit}}}{\nu_e} \right)^{1/2} \right\} \Leftrightarrow \\ (-c_1 X^{1/2} \leq Z) &\cap (Z \leq c_1 X^{1/2}) = V_1 \cap V_2. \end{aligned} \tag{A.4}$$

Also, equation (A.2) can be rewritten as

$$\begin{aligned} 2\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} &\leq \delta \Leftrightarrow \\ \frac{\hat{\sigma}_*}{\sigma_*} &\leq \frac{\delta}{2\sigma_*(f_{\text{crit}} m)^{1/2}} \Leftrightarrow \\ \left(\frac{\hat{\sigma}_*}{\sigma_*} \right)^2 \nu_e &\leq \nu_e \left\{ \frac{\delta}{2\sigma_*(f_{\text{crit}} m)^{1/2}} \right\}^2 \Leftrightarrow \\ X &\leq x_1. \end{aligned} \tag{A.5}$$

In turn, (A.3) can be rewritten as

$$\begin{aligned} \{ \hat{\theta}_d < -\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \} &\cup \{ \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} < \hat{\theta}_d \} \Leftrightarrow \\ \left\{ \frac{\hat{\theta}_d - \theta_d}{(\sigma_*^2 m)^{1/2}} < \frac{-\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} - \theta_d}{(\sigma_*^2 m)^{1/2}} \right\} &\cup \left\{ \frac{\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} - \theta_d}{(\sigma_*^2 m)^{1/2}} < \frac{\hat{\theta}_d - \theta_d}{(\sigma_*^2 m)^{1/2}} \right\} \Leftrightarrow \\ \left\{ \frac{\hat{\theta} - \theta}{(\sigma_*^2 m)^{1/2}} < \frac{-\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} - \theta_d}{(\sigma_*^2 m)^{1/2}} \right\} &\cup \left\{ \frac{\hat{\sigma}_*(f_{\text{crit}} m)^{1/2} - \theta_d}{(\sigma_*^2 m)^{1/2}} < \frac{\hat{\theta} - \theta}{(\sigma_*^2 m)^{1/2}} \right\} \Leftrightarrow \\ \left\{ Z < -\left(\frac{X f_{\text{crit}}}{\nu_e} \right)^{1/2} - \frac{\theta_d}{(\sigma_*^2 m)^{1/2}} \right\} &\cup \left\{ \left(\frac{X f_{\text{crit}}}{\nu_e} \right)^{1/2} - \frac{\theta_d}{(\sigma_*^2 m)^{1/2}} < Z \right\} \Leftrightarrow \\ (Z < -c_1 X^{1/2} - c_2) &\cup (c_1 X^{1/2} - c_2 < Z) = R_1 \cup R_2. \end{aligned} \tag{A.6}$$

We know $Z = (\hat{\theta} - \theta)/(\sigma_*^2 m)^{1/2} \sim \mathcal{N}(0, 1)$. Then, $Z/(X/\nu_e)^{1/2} \sim t(\nu_e)$, so that $Z/(X/\nu_e)^{1/2}$ follows a central t distribution with ν_e d.f. and Z and X are independent. Since $\Pr \{ V \} = 1 - \alpha$, computing $\Pr \{ (W \cap R) | V \}$ reduces to

considering $\Pr\{W \cap R \cap V\}$, namely,

$$\begin{aligned} \Pr\{(X \leq x_1) \cap (R_1 \cup R_2) \cap (V_1 \cap V_2)\} &= \\ &= \int_0^{x_1} \Pr\{(R_1 \cup R_2) \cap (V_1 \cap V_2) | (X = x)\} f_{\chi^2}(x; \nu_e) dx \\ &= \int_0^{x_1} \Pr\{(V_1 \cap V_2 \cap R_1) \cup (V_1 \cap V_2 \cap R_2) | (X = x)\} \\ &\qquad\qquad\qquad f_{\chi^2}(x; \nu_e) dx \\ &= \int_0^{x_1} \Pr\{\emptyset \cup (V_1 \cap V_2 \cap R_2) | (X = x)\} f_{\chi^2}(x; \nu_e) dx \\ &= \int_0^{x_1} \Pr\{[\max(c_1 X^{1/2} - c_2, -c_1 X^{1/2}) < Z \leq c_1 X^{1/2}] | \\ &\qquad\qquad\qquad (X = x)\} f_{\chi^2}(x; \nu_e) dx \\ &= \int_0^{x_1} [\Phi(c_1 x^{1/2}) - \Phi\{\max(c_1 x^{1/2} - c_2, -c_1 x^{1/2})\}] \\ &\qquad\qquad\qquad f_{\chi^2}(x; \nu_e) dx, \end{aligned} \tag{A.7}$$

with $c_2 > 0$ implying $(V_1 \cap V_2 \cap R_1) = \emptyset$.

Proof of Corollary 1. By changing the definition of rejection to $\theta_0 < L$ in the proof of the Theorem, R_1 is eliminated and the Corollary 1 result follows directly.

Proof of Corollary 2. Define a lower one-sided $100(1 - \alpha)\%$ confidence interval for θ as

$$\begin{aligned} 1 - \alpha &= \Pr\left\{-\infty < \frac{\hat{\theta} - \theta}{(\hat{\sigma}_*^2 m)^{1/2}} \leq f_{\text{crit}}^{1/2}\right\} \\ &= \Pr\{\hat{\theta} - \hat{\sigma}_*(f_{\text{crit}} m)^{1/2} \leq \theta < \infty\} \\ &= \Pr\{L \leq \theta\} \\ &= \Pr\{V\}. \end{aligned} \tag{A.8}$$

Eliminating the event V_1 in addition to R_1 in the proof of the Theorem leads immediately to the Corollary 2 result. Note that width, more accurately the lower half-width, is defined as $\hat{\theta} - L \leq \delta_L$. If δ_L equals $\delta/2$, then $\Pr\{U - L \leq \delta\} = \Pr\{\hat{\theta} - L \leq \delta_L\}$ and the width criterion has the identical effect on sample size in the two situations.

Proof of Corollary 3. By changing the definition of rejection to $U < \theta_0$ in the Theorem proof, R_2 is eliminated and the Corollary 3 result follows directly.

Proof of Corollary 4. Define an upper one-sided $100 \times (1 - \alpha)\%$ confidence interval for θ as

$$\begin{aligned} 1 - \alpha &= \Pr\left\{-(f_{\text{crit}})^{1/2} \leq \frac{\hat{\theta} - \theta}{(\hat{\sigma}_*^2 m)^{1/2}} < \infty\right\} \\ &= \Pr\{-\infty < \theta \leq \hat{\theta} + \hat{\sigma}_*(f_{\text{crit}} m)^{1/2}\} \\ &= \Pr\{\theta \leq U\} \\ &= \Pr\{V\}. \end{aligned} \tag{A.9}$$

Eliminating the event V_2 in addition to R_2 in the proof of the Theorem leads immediately to the Corollary 4 result. Note that width, more accurately described as the upper half-width, is defined as $U - \hat{\theta} \leq \delta_U$. If δ_U equals $\delta/2$, then $\Pr\{U - L \leq \delta\} = \Pr\{U - \hat{\theta} \leq \delta_U\}$ and the width criterion has the identical effect on sample size in the two situations.

Proof of Corollary 5. The results for each case of $\Pr\{W | V\}$ can be obtained immediately from the proofs of the Theorem and Corollaries 1 through 4 by eliminating the event rejection.

