

In press: Statistics in Medicine

Noninferiority trial designs for odds ratios and risk differences

Joan F. Hilton

University of California San Francisco

SUMMARY

This study presents constrained maximum likelihood derivations of the design parameters of noninferiority trials for binary outcomes with the margin defined on the odds ratio (ψ) or risk difference (δ) scale. The derivations show that, for trials in which the group-specific response rates are equal under the point-alternative hypothesis, the common response rate, π^N , is a fixed design parameter whose value lies between the control and experimental rates hypothesized at the point-null, $\{\pi_C, \pi_E\}$. We show that setting π^N equal to the value of π_C that holds under H_0 underestimates the overall sample size requirement. Given $\{\pi_C, \psi\}$ or $\{\pi_C, \delta\}$ and the type I and II error rates, an algorithm finds clinically meaningful design values of π^N , and the corresponding minimum asymptotic sample size, $N = n_E + n_C$, and optimal allocation ratio, $\gamma = n_E/n_C$. We find that optimal allocations are increasingly imbalanced as ψ increases, with $\gamma_\psi < 1$ and $\gamma_\delta \approx 1/\gamma_\psi$, and that ranges of allocation ratios map to the minimum sample size. The latter characteristic allows trialists to consider trade-offs between optimal allocation at a smaller N and a preferred allocation at a larger N . For designs with relatively large margins (e.g., $\psi > 2.5$), trial results that are presented on both scales will differ in power, with more power lost if the study is designed on the risk difference scale and reported on the odds ratio scale than vice versa.

KEY WORDS: active-controlled trial, allocation ratio, ancillary parameter

Correspondence to:

Joan F. Hilton, Sc.D., M.P.H.

Department of Epidemiology & Biostatistics

University of California San Francisco

185 Berry Street, Suite 5700

San Francisco CA 94107-1762 USA

e-mail: joan@biostat.ucsf.edu phone: 415-514-8029 fax: 415-514-8150

1 INTRODUCTION

A search of the PubMed database conducted on Midsummer’s Day 2009, using criteria “New Engl J Med [JO] AND noninferiority,” provided insights into how the noninferiority trial design is being applied in medical studies. The search identified 32 papers, published 2001-2009, each describing a distinct randomized controlled trial. Only one trial studied a continuous primary outcome [1], 13 studied time-to-event outcomes [2-14], and 21 studied binary outcomes [12-32], demonstrating the relative importance of these outcomes in medical clinical trials. The designs of the 21 trials with binary outcomes can be characterized as follows.

- Seventeen trials expressed the noninferiority margin as a risk difference, two as odds ratios [13,15], and two as a relative risks [16,17] – apparently reflecting trialists’ preferences. All trial designs specified equivalent experimental (E) and control (C) response rates under the alternative hypothesis, H_A .
- Extreme response rates were more common than moderate response rates. Six papers seemed to specify a rate assuming equality of the experimental and control groups [12,15, 18-21]; seven seemed to specify the control group response rate, π_C , assuming inequality [14,15,17,21-25]; and ten did not cite a design rate – in these cases we used the empirical estimate of π_C from the paper’s *Results* section to classify extremity of the rate. Based on this mixture of definitions, the “response rate,” π , was within .10 of either 0 or 1 for eight trials (38%), within .20 for 13 trials (62%), and within .30 for 18 trials (86%).
- All trials appeared to rely on the normal approximation to the binomial distribution to generate the overall sample size, $N = n_C + n_E$, since none indicated use of an exact distribution. This assumption was generally reasonable as only three “expected values,” $N\pi$, fell below 30 $\{N=183, \pi=.06$ [20]; $N=340, \pi=.06$ [26]; $N=667, \pi=.012$ [27]}.
- Although imbalanced allocation can reduce the sample size requirement of noninferiority trials, this advantage was rarely employed; only two trials used imbalanced allocation to groups [12,27].

Several authors have discussed imbalanced allocation ratios as optimal for noninferiority trials [33-35]. Among these, De Boo and Zielhuis [33] studied designs with $.01 \leq \pi_C \leq .10$, focusing on

failure rates. Their results show that designs based on the risk difference call for allocating more patients to the experimental group ($n_E > n_C$) whereas those based on the relative risk call for the opposite ($n_E < n_C$). They provided precise sample sizes for combinations of π_C and noninferiority margins. Considering that randomized trials often achieve the target sample size only approximately – due to over-running, loss to follow-up, or use of the per-protocol rather than the intent-to-treat sample – we examined how important it is to precisely achieve a target allocation ratio.

The current study examines designs of noninferiority trials for binary outcomes with two objectives: (i) to clarify the role of the common response rate, π^N , under the alternative hypothesis that the experimental and control rates are equal, and to offer strategies for specifying its value, and (ii) to explore the relationship between the scale of the noninferiority margin – we consider risk differences and odds ratios – and the overall sample size, allocation ratio, and power. We examine the setting where n_E and n_C patients are randomly allocated to experimental and control therapies, respectively, in the ratio $\gamma = n_E/n_C$, and the numbers of responders are binomially distributed, $y_E \sim Bi(n_E, \pi_E)$ and $y_C \sim Bi(n_C, \pi_C)$. Under the point-null hypothesis, H_0 , we assume that the response rates differ by the noninferiority margin, specified by $\delta = \pi_E - \pi_C$ or $\log \psi = \log(\pi_E/(1 - \pi_E)) - \log(\pi_C/(1 - \pi_C))$. Maximum likelihood (ML) estimates of these parameter values are given by $\hat{\delta} = \hat{\pi}_E - \hat{\pi}_C$ and $\log \hat{\psi} = \log[\hat{\pi}_E/(1 - \hat{\pi}_E)] - \log[\hat{\pi}_C/(1 - \hat{\pi}_C)]$, respectively, where $\hat{\pi}_j = y_j/n_j$, $j = C, E$, estimate the response rates. Under the point-alternative hypothesis of equal response rates, we denote the common response rate by π^N . As illustrated in this paragraph, in this paper we use π_C to refer to the value of the control-group response rate when it doesn't equal π_E (i.e., under H_0 ; typically derived from subject-matter literature) and π^N to refer to its value when these parameters are assumed to be equal (i.e., under H_A). In contrast, under H_A , DeBoo and Zielhuis [33] denote the common value by π_C and Farrington and Manning [34] and Miettinen and Nurminen [36] provide the values of both π_C and π_E to show they are equal.

To further set the stage for the current study, we examine Shiffman et al [15] more closely. This noninferiority trial hypothesized that short-term therapy (16 weeks; experimental arm) and long-term therapy (24 weeks; control arm) for hepatitis C virus (HCV) infections have similar efficacy, as measured by rates of undetectable HCV RNA 24 weeks after the end of therapy [15].

To design the trial, values of the control group success rate and the noninferiority margin on the risk difference scale were specified; the authors then converted δ to ψ for sample size calculation. They planned to enroll 700 patients per group in order to have 80% power using a two-sided 95% confidence interval on the odds ratio. In Section 4.3, we will discuss the authors' choice of balanced allocation, $\gamma = 1$; their use of π_C versus π^N as a design parameter; and whether one can base the design on one scale of the noninferiority margin and inferences on the other scale without loss of power (this was not done by Shiffman et al [15] but was done by others).

In Section 2 we present the noninferiority hypotheses more formally and present the equation for asymptotic sample size calculation that we will use. In Section 3, we use maximum likelihood derivations to show that π^N is a fixed design parameter for noninferiority trials that assume equal response rates under H_A , and that its value does not equal π_C . In addition, we show that the expression for the optimal allocation ratio depends on the values of the design parameters and differs according to the scale of the noninferiority margin, for δ and $\log \psi$. Because π^N is unknown at the design stage of a trial, in Section 4 we propose an approach to specifying its value and we present an algorithm for finding π^N , N and γ . We apply these methods to the design of Shiffman et al [15], and then to a wide range of design parameters. We conclude in Section 5 with a summary of these findings.

2 NONINFERIORITY TRIAL DESIGN

2.1 *The hypotheses and the noninferiority margin*

Let δ and $\log \psi$ be members of a family of parameters, θ , that contrast two groups' response rates; hence, θ represents δ and $\log \psi$ more generally. Let θ_0 and θ_A be the values of θ at the point-null and point-alternative hypotheses. Here, we restrict attention to $\theta_A = 0$ (that is, to equality of response rates) as the clinically most relevant choice. The quantity θ_0 , referred to as the noninferiority margin, is the smallest $\pi_E:\pi_C$ contrast that is believed to be consistent with clinically acceptable inferiority (i.e., noninferiority). We define the contrasts so that $\theta_0 > 0$.

The noninferiority hypotheses can be understood in terms of either within-group parameters (here, failure rates or odds of failure) or between-group parameters (risk differences or log odds ratios); the latter may be centered or not. When response rates represent failures, the response

rate scale illustrates our expectation under H_A that $\pi_E < \pi_C + \theta_0$, as well as our assumption that the failure rates on both active treatments are lower than the putative failure rate on placebo, π_P (Figure 1(a)). However, it is on the centered-contrast scale, $\theta - \theta_0$, that sample size and power are calculated and the hypothesis test is conducted. On this scale, if the confidence interval on $\hat{\theta} - \theta_0$ lies below 0 we conclude that E is noninferior to C , and if it lies below $\theta_A - \theta_0$ we conclude that E is superior to C (Figure 1(b)). On this scale, the hypotheses are:

$$H_0 : \theta - \theta_0 \geq 0 \text{ versus } H_A : \theta - \theta_0 < 0.$$

<< Figure 1 about here. >>

2.2 Sample size and power formulae

The asymptotic sample size requirement corresponding with pre-specified nominal size and power levels, α and $1 - \beta$, is defined to detect the centered noninferiority margin, $\theta_A - \theta_0$, using test statistic

$$T(\hat{\theta}) = \sqrt{n_C}(\hat{\theta} - \theta_0)/\tilde{\sigma}_0(\hat{\theta}), \quad (1)$$

where $\tilde{\sigma}_0(\hat{\theta})$ is based on $\tilde{\pi}_C$ estimated at the analysis stage under the point-null hypothesis constraint, $\tilde{\pi}_E - \tilde{\pi}_C = \theta_0$ (Miettinen and Nurminen [36]). Farrington and Manning [34] found the asymptotic control-group sample size requirement corresponding with this test statistic to be

$$n_C \geq \frac{\left\{ z_{1-\alpha} \tilde{\sigma}_0(\hat{\theta}) + z_{1-\beta} \tilde{\sigma}_A(\hat{\theta}) \right\}^2}{(\theta_A - \theta_0)^2}, \quad (2)$$

where $z_{1-\epsilon}$ is a critical value of the standard Normal distribution such that $\Pr\{Z < z_{1-\epsilon}\} = \epsilon$, and $\tilde{\sigma}_0(\hat{\theta})$ is a function of a large-sample approximation of $\tilde{\pi}_C$ estimated at the design stage under the point-null constraint, $\tilde{\pi}_E - \tilde{\pi}_C = \theta_0$ (see Section 3). Scaling n_C by the allocation ratio, $\gamma = n_E/n_C$, yields the overall sample size,

$$N = \text{round}\{(1 + \gamma)n_C\}. \quad (3)$$

Subsequently, we round n_C and calculate $n_E = N - n_C$. The power associated with the given design is

$$1 - \beta = \Phi^{-1}\{[|\theta_A - \theta_0|\sqrt{n_C} - z_{1-\alpha} \bar{\sigma}_0(\hat{\theta})]/\bar{\sigma}_A(\hat{\theta})\}, \quad (4)$$

while the $100(1 - \alpha/2)\%$ confidence interval under H_A is $\hat{\theta} \pm z_{1-\alpha/2}\bar{\sigma}_A(\hat{\theta})$.

Some authors reduce the numerator of equation (??) to $(z_{1-\alpha} + z_{1-\beta})\bar{\sigma}_A(\hat{\theta})$ [37, 38]. We recommend equation (??) because differing values of $\bar{\sigma}_0(\hat{\theta})$ and $\bar{\sigma}_A(\hat{\theta})$ could substantially alter n_C . In the next section, we derive expressions for $\bar{\sigma}_0(\hat{\theta})$, $\bar{\sigma}_A(\hat{\theta})$, and γ_θ from the likelihood function. Our derivation for $\theta = \delta$ (see Appendix) yields sample sizes identical to those calculated by Farrington and Manning [34] at pre-specified values of γ ; whereas those authors also address the relative risk, we also address $\theta = \log \psi$. Roebuck and Kühn [39] compared Farrington and Manning’s [34] constrained maximum likelihood approach with other asymptotic methods over a wide range of design parameters and reported excellent performance. De Boo and Zielhuis [33] compared Farrington and Manning’s [34] approach with exact unconditional inference and recommended the latter when “some elevation ... of the type I error is acceptable.”

3 π^N IS A FIXED DESIGN PARAMETER FOR NONINFERIORITY TRIALS

Among confusions about the noninferiority trial design is the role of π^N . For superiority trial designs, π_C is a fixed design parameter; perhaps investigators are extending that role when they specify it as a design parameter for noninferiority trials. However, the derivations in this section and the appendix examine the theory behind the choice and show that π^N is the appropriate fixed design parameter for noninferiority trials when $\theta_A = 0$.

3.1 *Margin parameterized by the log-odds ratio, $\theta = \log \psi$*

The noninferiority likelihood function under H_0 is parameterized by π_C and ψ . In turn, the score equation is a function of the ML estimate of the control-group response rate, $\tilde{\pi}_C$, defined under the null hypothesis constraint, $\text{logit}(\tilde{\pi}_E) - \text{logit}(\tilde{\pi}_C) - \log \Psi = 0$:

$$(y_C - n_C\tilde{\pi}_C) + (y_E - n_E\tilde{\pi}_E) = 0,$$

$$\text{where } \tilde{\pi}_E = \psi\tilde{\pi}_C/[1 - (1 - \psi)\tilde{\pi}_C].$$

Its expectation at the point-alternative shows that the limiting value of the constrained ML estimate, $\vec{\pi}_C = \lim_{H_A: N \rightarrow \infty} \tilde{\pi}_C$, depends on three terms: π^N , the common value of π_C and π_E under H_A ; ψ , the noninferiority margin; and γ , the allocation ratio:

$$\pi^N = (1 - \omega_\psi) \vec{\pi}_C + \omega_\psi \vec{\pi}_E, \quad (5)$$

where

$$\begin{aligned} \vec{\pi}_E &= \psi \vec{\pi}_C / [1 - (1 - \psi) \vec{\pi}_C], \\ \omega_\psi &= \gamma_\psi / (\gamma_\psi + 1), \quad \text{and} \\ \gamma_\psi &= \frac{\pi^N - \vec{\pi}_C}{\vec{\pi}_E - \pi^N}. \end{aligned}$$

From equation (??) and its components, one can see that $\{\pi^N, \psi\}$ are parameter values, not estimates, and thus they must be specified by the design. The objective, then, is to find the solution-pair $\{\gamma, \vec{\pi}_C\}$ associated with the minimum N , which cannot be found in closed form. Since N depends on the standard deviations, these terms also must be defined. Using the delta method, if $u = \hat{\pi}$ and $v = \text{logit}(\hat{\pi})$, then $\text{var}[\text{logit}(\hat{\pi})] = [\pi(1 - \pi)]^{-2} \pi(1 - \pi)/k = [k \pi(1 - \pi)]^{-1}$. Further, given independent groups, the limiting standard deviations of $\log \hat{\psi}$ are [35]

$$\begin{aligned} \vec{\sigma}_0(\log \hat{\psi}) &= \left\{ \frac{1}{\vec{\pi}_C(1 - \vec{\pi}_C)} + \frac{1}{\vec{\pi}_E(1 - \vec{\pi}_E)\gamma_\psi} \right\}^{0.5}, \\ \vec{\sigma}_A(\log \hat{\psi}) &= \left\{ \frac{1}{\pi^N(1 - \pi^N)} \left(1 + \frac{1}{\gamma_\psi} \right) \right\}^{0.5}. \end{aligned}$$

3.2 Margin parameterized by the risk difference, $\theta = \delta$

By an analogous process, when the likelihood is parameterized by π_C and δ , the score equation gives rise to an ML estimate, $\tilde{\pi}_C$, defined under the the null hypothesis constraint $\tilde{\pi}_E - \tilde{\pi}_C - \delta = 0$:

$$\frac{y_C - n_C \tilde{\pi}_C}{\tilde{\pi}_C(1 - \tilde{\pi}_C)} + \frac{y_E - n_E \tilde{\pi}_E}{\tilde{\pi}_E(1 - \tilde{\pi}_E)} = 0. \quad (6)$$

The standard deviation of $\hat{\delta}$ at θ_0 , can be found by taking the derivative of the score equation, inverting its negative expectation under H_0 , and taking the square root (see Appendix):

$$\tilde{\sigma}_0(\hat{\delta}) = \{\tilde{\pi}_C(1 - \tilde{\pi}_C) + \tilde{\pi}_E(1 - \tilde{\pi}_E)/\gamma\}^{0.5}. \quad (7)$$

However, because sample size and power calculations assume that H_A holds, equations (??) and (??) must be based on the limiting value of $\tilde{\sigma}_0(\hat{\delta})$ at the point-alternative hypothesis (see Appendix). This term is obtained by expressing equation (??) as a function of $\bar{\pi}_C = \lim_{H_A: N \rightarrow \infty} \tilde{\pi}_C$, which is found by taking the expectation of equation (??) at θ_A :

$$\pi^N = (1 - \omega_\delta) \bar{\pi}_C + \omega_\delta \bar{\pi}_E, \quad (8)$$

where

$$\begin{aligned} \bar{\pi}_E &= \bar{\pi}_C + \delta, \\ \omega_\delta &= \frac{\gamma_\delta}{\gamma_\delta + R}, \quad \text{for } R = \frac{\bar{\pi}_E(1 - \bar{\pi}_E)}{\bar{\pi}_C(1 - \bar{\pi}_C)}, \quad \text{and} \\ \gamma_\delta &= \left(\frac{\pi - \bar{\pi}_C}{\bar{\pi}_E - \pi} \right) R. \end{aligned}$$

As when ψ was the noninferiority margin, the optimal value of γ – that which minimizes N – must be found iteratively. When δ is the noninferiority margin the limiting standard deviations of $\hat{\delta}$ are [39; Appendix]

$$\bar{\sigma}_0(\hat{\delta}) = \{\bar{\pi}_C(1 - \bar{\pi}_C) + \bar{\pi}_E(1 - \bar{\pi}_E)/\gamma_\delta\}^{0.5}, \quad (9)$$

$$\bar{\sigma}_A(\hat{\delta}) = \left\{ \pi^N(1 - \pi^N) \left(1 + \frac{1}{\gamma_\delta} \right) \right\}^{0.5}. \quad (10)$$

Equations (??) and (??) show that one can solve for a unique value of $\bar{\pi}_C$, given $\{\pi^N, \theta\}$ and γ , and that the values of $\bar{\pi}_C$ used in a sample size calculation can differ between parameterizations, θ .

4 ALGORITHMS TO IDENTIFY THE OPTIMAL γ AND MINIMUM N

Any investigator-selected value of γ_θ can be used in equations (??)–(??), including the balanced case, $\gamma_\theta = 1$. However, since sub-optimal choices yield excessive sample sizes and inadequate power, it would be useful to know the $N \times \gamma$ relationship at the design stage so that informed choices could be made. We created two algorithms to find these values: The *Fixed- π^N Algorithm*, which assumes a known value of π^N , and the *Fixed- π_C Algorithm*, which assumes a known value of π_C but an unknown value of π^N (see <http://www.epibiostat.ucsf.edu/biostat/joan/>).

4.1 Finding $\{N, \gamma\}$ assuming π^N is known: The *Fixed- π^N Algorithm*

Given $\{\pi^N, \delta, \alpha, 1 - \beta\}$, the algorithm loops through candidate values of $\tilde{\pi}_C$ ranging from $\max(\pi^N - \delta, .0001)$ to π^N , incremented by .0001, and generates the corresponding value of γ and $\log \psi = \text{logit}(\tilde{\pi}_C + \delta) - \text{logit}(\tilde{\pi}_C)$ for each; only $\gamma \in (.25, 4.0)$ are retained. Among these candidate pairs, the optimal solutions $\{\gamma_\theta, \tilde{\pi}_C\}$ are those associated with N_θ , the minimum value of N .

Figure 2 displays the algorithm's results for nine $\pi^N \times \gamma$ combinations studied by Farrington and Manning [34] when $\theta = \delta$, for $\delta = .20$, and $\{\alpha, 1 - \beta\} = \{.05, .90\}$: $\pi^N = .10, .05$, and $.01$ at $\gamma = 1.5, 1.0$, and $.67$. The *Fixed- π^N Algorithm* essentially confirms their results (see table). For example, when $\pi^N = .10$ and $\gamma = 1.5, 1.0$, and $.67$, we find $N = 85, 92$, and 105 , while they find $N = 86, 92$, and 105 , respectively. However, our algorithm finds that $\gamma = 2.5$ and 3.5 also yield $N = 85$ and 92 , respectively. Thus it reveals the nonuniqueness of the $N \times \gamma$ relationship, a feature that could be leveraged in a particular study design application. Rather than specify γ a priori and identify the corresponding N , our algorithm identifies the smallest overall sample size requirement and the corresponding range of allocation ratios; for $\pi^N = .10$, $N_\delta = 83$ and $\gamma_\delta \in (1.88, 2.09)$ (i.e., $n_E/n_C = 54/29, 55/28, 56/27$). As the sawtoothed $N \times \log \gamma$ plots of Figure 2 suggest, slightly larger values of N lie in this same range as well. For example, $N = 85$ is associated with $\gamma_\delta \in (1.43, 2.71)$, occurring at each value of n_C from 35 to 23. However, because of the concave relationship consistent with ML estimation, $N = 92$ is associated with two widely disparate sets of solutions with $0.25 < \gamma < 4.0$: $\gamma_\delta \in (0.96, 1.04)$ and $\gamma_\delta \in (3.38, 3.84)$.

<< Figure 2 about here. >>

4.2 Finding $\{N, \gamma\}$ when π^N is unknown: The Fixed- π_C Algorithm

In practice, information typically is available about the control-group response rate but not about the common response rate, π^N . We defined the *Fixed- π_C Algorithm* to examine two approaches that investigators are likely to use to obtain a value of π^N from the known value of π_C .

Investigators might use the Midpoint Approach because of its simplicity and its familiarity in the setting of superiority trials under balanced allocation (see [41]). Alternatively, the Tailored Approach allows one to select π^N so that $\vec{\pi}_C$ is consistent with the known value of π_C .

The Midpoint Approach specifies the common response rate at the midpoint of the group-specific response rates, $\pi^N = .5(\pi_E + \pi_C)$, where the definition of π_E is specific to θ (see equations (??) and (??)). Then, using the *Fixed- π^N Algorithm* with design pair $\{\pi^N, \theta\}$, it finds solution pairs $\{\gamma_\theta, \vec{\pi}_C\}$ associated with N_θ . For the setting of Figure 2, if $\pi_C = .05$ then $\pi^N = .15$ (not tabled). Given $\{\pi^N, \delta\} = \{.15, .20\}$, the algorithm finds $N_\delta = 111$ at $\{\gamma, \vec{\pi}_C\} \in \{1.42, .081\}, \dots, \{2.01, .070\}$. Note that this range of values of $\vec{\pi}_C$ excludes the known value, $\pi_C = .05$.

The Tailored Approach solves for $\{N, \gamma\}$ at the ML solution $\vec{\pi}_C$ that most closely matches the known value of π_C (e.g., that found in the subject-matter literature). The algorithm processes a series of designs indexed by π^N . For each $\pi^N \in (\pi_C, \pi_C + \delta)$ incremented by .001, it finds solution-pairs $\{\gamma_\theta, \vec{\pi}_C\}$ associated with N_θ . Among these solutions, it selects that for which the median value of $\vec{\pi}_C$ is closest to the known parameter π_C . For the setting of Figure 2, if $\pi_C = .05$ then $\pi^N = .115$. Given $\{\pi^N, \delta\} = \{.115, .20\}$, the algorithm finds $N_\delta = 92$ at $\{\gamma, \vec{\pi}_C\} \in \{1.57, .055\}, \dots, \{2.18, .046\}$. Note that this range of values of $\vec{\pi}_C$ includes the known value, $\pi_C = .05$.

4.3 Hepatitis C Virus Example.

To replicate the noninferiority trial design of Shiffman et al [15], where $\{\alpha, 1 - \beta\} = \{.025, .80\}$ and $\delta = .06$, we first adopt the position that, by “[we] assumed a sustained virologic response rate of 80% in both groups,” the authors meant to design the trial with (failure rate) $\pi^N = .20$. Using the *Fixed- π^N Algorithm*, on the risk-difference scale we find $N_\delta = 1,393$ at $\gamma_\delta \in (1.12, 1.20)$ where $\vec{\pi}_C \in (.172, .171)$ (Figure 3, top-right). On the log ψ scale, we define $\psi = 1.456$ at $\vec{\pi}_C = .171$, the

median of the range of $\vec{\pi}_C$ found above. We confirm the authors' calculation of $N_\psi = 1,399$ but find that this occurs at $\gamma_\psi \in (.813, .876)$, $\vec{\pi}_C \in (.173, .172)$, and $\delta(\vec{\pi}_C) = .0603$ (Figure 3, top-left).

For comparison, we adopt the position that the authors meant to design the trial “assuming a sustained virologic response rate of 80% in the 24-week group” with $\pi_C = .20$ and we select values of π^N via both *Fixed- π_C Algorithm* approaches outlined in Section 4.2. On the $\log \psi$ scale, we calculate $\psi = 1.405$ at $\pi_C \equiv .20$ and $\delta = .06$. Regardless of the contrast parameter or the approach used to select π^N , by definition (since $\vec{\pi}_C < \pi^N$) the *Fixed- π_C Algorithm* response rates are closer to .50 when $\vec{\pi}_C = .20$ than when $\pi^N = .20$; in turn, these drive up N and are associated with more balanced allocation ratios (Figure 3, top table). On the $\log \psi$ scale, the Midpoint Approach yields a slightly smaller N than the Tailored Approach, whereas on the δ scale, the opposite is true. Thus the approach used to select π^N affects N differently on the two parameter scales of interest. Since the Midpoint definition is arbitrary and the Tailored definition is clinically meaningful, we recommend using the latter approach only.

Whether the value of π^N is known or unknown, optimal allocation ratios on the risk-difference scale call for more patients on the experimental arm while those on the log-odds scale call for more patients on the control arm (Figure 3; see table). The inverse relationship of the rate-based terms of $\vec{\sigma}_A(\log \hat{\psi})$ and $\vec{\sigma}_A(\hat{\delta})$ (and $\vec{\sigma}_0(\log \hat{\psi})$ and $\vec{\sigma}_0(\hat{\delta})$) provides insight into this. More precisely, recall that the allocation ratio ties the location of π^N between the group-specific response rates to the balance between the sample sizes, n_E/n_C , and does so differently for these two contrast parameters. From equations (??) and (??), respectively,

$$\pi^N = (1 - \omega_\theta) \vec{\pi}_C + \omega_\theta \vec{\pi}_E,$$

where

$$\omega_\theta = \begin{cases} \gamma_\psi / (\gamma_\psi + 1), & \text{if } \theta = \log \psi, \\ \gamma_\delta / (\gamma_\delta + R), \text{ where } R = [\vec{\pi}_E(1 - \vec{\pi}_E)] / [\vec{\pi}_C(1 - \vec{\pi}_C)], & \text{if } \theta = \delta. \end{cases}$$

For the hepatitis C virus example, one can see that ω_ψ and ω_δ must be similar because $\vec{\pi}_C$ and

π^N are similar across parameterizations, regardless of whether π^N or π_C is fixed (Figure 3; see table); thus the differing values of γ_ψ and γ_δ arise primarily via $R > 1$ for this example.

Interestingly, although n_C and γ differ across parameterizations in this example, both combinations yield nearly identical overall sample sizes N (equation (??)). For other design parameters $\{\pi^N, \theta\}$, ω_ψ and ω_δ can differ as well as R (e.g., $\{.05, .06\}$ in Figure 3 (table)).

When $\pi^N = .20$, $\delta = .06$, and $\psi = 1.456$, despite that the optimal γ_ψ and γ_δ are nearly inverses, swapping parameterizations between the design and analysis stages of the trial causes negligible loss of power. Specifically, plugging all of the design values for the log-odds scale into equation (??), but applying the effect size and standard deviation formulas for the risk-difference scale, maintains 80% power. However, the reverse strategy maintains only 72% power – in part because N_δ is less than that required on the log-odds scale ($N_\psi = 1,399$). Furthermore, when $\pi^N = .10$ and $.05$, designs based on ψ that are analyzed on the δ scale maintain 78.8% and 74.6% power, respectively, but designs based on δ that are analyzed on the log ψ scale maintain only 76.1% and 62.8% power. Side-by-side plots of $N \times \log \gamma$, with matching design parameters but differing scales (Figure 3), suggest that as π^N decreases power is lost if the parameter scale is switched *between* the design and the analysis because $\{\gamma_\psi, \bar{\pi}_C(\psi)\}$ and $\{\gamma_\delta, \bar{\pi}_C(\delta)\}$ diverge. However, translating the noninferiority margin from the risk-difference to the log-odds scale *within* the design process, as Shiffman et al [15] did, does not harm the study design and can enhance understanding. (As an aside, since $\delta > 0$, we believe they inverted ψ in specifying $\psi = .70$.)

<< Figure 3 about here. >>

4.4 General patterns in N , γ , and power

Designs based on the risk-difference scale, $\theta = \delta$. To expand on the example above, we examined patterns in N_δ and the median values of γ and $\bar{\pi}_C$, as functions of design parameters $\{\pi^N, \delta\}$, when the one-sided $\alpha = .025$ and power is 80%. For $\pi^N < .50$ and $\delta \leq .20$, $\gamma_\delta \geq 1$, in general, and grows more balanced as π^N nears $.50$ and as the noninferiority margin decreases (Table 1.a).

Solving for ψ at the median value of $\bar{\pi}_C$ shows that $\psi(\bar{\pi}_C)$ (i.e., ψ defined at $\bar{\pi}_C$, given δ) decreases in π^N and increases in δ such that our $\pi^N \times \delta$ combinations with $\delta > \pi^N$ yield $\psi > 4$

(Table 1.b). Further, when we calculated the power associated with a design on the risk-difference scale but an analysis on the log-odds scale, $\pi^N \times \delta$ combinations with $\psi(\bar{\pi}_C) > 2.5$ had significantly reduced power and few expected control responses. Combinations with $\psi(\bar{\pi}_C) > 4$ had $< 50\%$ power and negligible responses in the control group (see Table 1.b).

Designs based on the log-odds scale, $\theta = \log \psi$. Using the values of $\{\pi^N, \psi\}$ from the analysis above as design parameters (Table 1.b), we examined patterns in N_ψ and the corresponding median values of γ and $\bar{\pi}_C$, when the one-sided $\alpha = .025$ and power is 80%. In order to present these findings by π^N and δ as above, we solved for values of δ at the median values of $\bar{\pi}_C$ and rounded them. Generally, $\delta(\bar{\pi}_C)$ (i.e., δ defined at $\bar{\pi}_C$, given ψ) is at least as large as the corresponding δ ; however, at the four largest values of ψ they are smaller (i.e., at $\psi > 20$; see Table 2.b, upper rows). Two are so much smaller that the corresponding data are not entered in Table 2.a: $\{N_\psi, \gamma_\psi, \bar{\pi}_C\} = \{633, .28, .0008\}$ for $\psi = 57$ and $= \{983, .35, .0003\}$ for $\psi = 130$.

We found that $N_\psi \approx N_\delta$ and $\gamma_\psi \approx 1/\gamma_\delta$, in general. However, for designs with $\delta > \pi^N$, (i) $N_\psi > N_\delta$, (ii) γ_ψ is less balanced than γ_δ , and (iii) the expected values in the control group are larger on the log-odds scale, which can be attributed to $\gamma_\psi < 1$ (i.e., $n_E < n_C$) as well as to larger values of $\bar{\pi}_C$. Clearly, significant mismatches between designs occur as δ grows large relative to π^N , which is when ψ grows large. Because N_ψ tends to be slightly larger than N_δ , the consequences for power are milder if the design is based on the log-odds but the analysis is based on the risk difference than the reverse: Even the four cases with $\delta > \pi^N$ maintain $\geq 65\%$ power, and all other cases (i.e., $\delta \leq \pi^N$) maintain 75%-80% power (Table 2.b, lower row).

5 Discussion

We present a constrained maximum likelihood derivation of the design parameters for noninferiority trials that differs from that in previous reports but produces similar results. Our derivation reveals that when π^N is a design parameter for noninferiority trials (that is, when the design assumes equal response rates under H_A), the value of π^N under H_A does not equal the control-group response rate under H_0 , π_C ; rather, π^N lies between the pair of group-specific response rates. By analogy, for the superiority trial design, π_C also “takes different values” under the point-null and point-alternative hypotheses. In the superiority setting, π_C is the fixed design

parameter under H_A and $\bar{\pi}^S = (1 - \omega)\pi_C + \omega\pi_E$ is estimated under H_0 ; in particular, $\bar{\pi}^S = .5(\pi_C + \pi_E)$ when $\omega = .5$. In both settings, the common value under equal response rates is a weighted average of the group-specific response rates [41].

Miettinen and Nurminen [36] presented a closed-form solution to $\bar{\pi}_C$, from which $\bar{\sigma}_0(\hat{\theta})$ can be calculated for use in equation (2). The inputs to their calculation are the margin, θ_0 ; the response rates under H_A , $\{\pi_C, \pi_E\}$, which, when equal, specify π^N ; and the design value of γ . Additional processing of their output to identify the minimum N and the corresponding pairs $\{\gamma, \bar{\pi}_C\}$ reveals slightly higher minimum sample sizes and narrower ranges of solution pairs than are found by our algorithm. This occurs because their algorithm smooths the sawtoothed $N \times \log \gamma$ relationship (Figure 2), such that only a single value of N corresponds with a range of values of γ .

Judging by the design information presented in 21 reports of noninferiority trials for binary outcomes that were recently published in *New England Journal of Medicine*, we believe that confusion about the study design roles of π^N and π_C is common. We show that equating π^N with the value of π_C anticipated under H_0 substantially underestimates the sample size by locating the design too low on the response-rate scale. Because the value of π^N is rarely known from preliminary data, we offer two approaches to specifying its value. We strongly recommend our Tailored Approach, which defines π^N such that the design-stage estimate, $\bar{\pi}_C$, is as close as possible to its known value. In turn, for communication with investigators, response rates and between-group contrasts can be plotted (Figure 1). We find it extremely useful to confirm that a design is clinically reasonable on both the risk-difference ($\theta = \delta$) and log-odds ($\theta = \log \psi$) scales. Our result show that, as a rough rule of thumb, the margin is excessively large when $\delta > \pi^N$ and results in negligible expected response counts in the control group. When an expected response count is small, exact inference is recommended [42]. Further discussions of clinically appropriate noninferiority margins are presented elsewhere [43-45].

Our review of published noninferiority trials shows that most trialists employ balanced allocation to groups in the binary-outcome noninferiority setting, despite that several authors have shown [33-35,41] and our results confirm that imbalanced allocation reduces the overall sample size requirement, especially as ψ increases. For ψ in the range 2.5 to 3.5, which coincide with $\pi^N \approx \delta$, we show that optimal allocations are $\gamma \in (1.67, 2.0)$ when the margin is defined on the risk-difference scale and approximately the inverse on the log-odds scale. We found that even

at these clinically reasonable values of the noninferiority margin, where γ_δ and γ_ψ are very different, little power is lost if the design is conducted on the log-odds scale and the analysis on the risk-difference scale but that the power loss can be substantial in the opposite direction. Three papers included in our review reported treatment contrasts and their 95% confidence intervals on two scales [12-14] – most likely assuming that both analyses have the same power. Following the example of Shiffman et al [15], authors more comfortable with expressing the noninferiority margin on the risk-difference scale could translate that value to the log-odds scale at the design stage, but not between the design and analysis stages.

We found that designs based on the log-odds scale have three important advantages. First, when π^N is known or is selected via our Tailored Approach, $N_\psi > N_\delta$; the larger sample size ensures adequate power for reporting results on either this scale or the risk-difference scale. Second, since $\gamma_\psi < 1$, the group-specific response rate – which is closer to the boundary of the parameter space – has the larger sample size, ensuring that both groups have adequate expected response counts (unless the margin is excessively large). Finally, noninferiority trials are typically analyzed in the per-protocol sample, in which some benefits of randomization may have been lost. To compensate for this, an adjusted analysis via logistic regression modeling can easily be conducted on the log-odds scale. A disadvantage of the log-odds scale, however, is that a less than 50:50 chance of allocation to the experimental treatment could slow accrual.

Regardless of the scale of the noninferiority margin, the range of γ associated with the minimum or near minimum N can be quite wide, offering trialists deliberate and conscious choice in the selection of an allocation ratio, once the optimal range is identified. Our algorithms obviate the practice of choosing N at a pre-specified value of γ .

APPENDIX 1

Details of derivation applicable to $\theta = \delta$; details for $\theta = \psi$ are analogous:

$$L_{H_0}(\pi_C|\delta) = [\pi_E^{x_E}(1 - \pi_E)^{n_E - x_E}][\pi_C^{x_C}(1 - \pi_C)^{n_C - x_C}], \quad \text{where } \pi_E = \pi_C + \delta$$

$$\log L_{H_0}(\pi_C|\delta) = x_E \log(\pi_C + \delta) + (n_E - x_E) \log[1 - (\pi_C + \delta)] + x_C \log(\pi_C) + n_C \log(1 - \pi_C)$$

$$\frac{\partial}{\partial \pi_C} \log L_{H_0}(\pi_C|\delta) = (x_E - n_E \pi_E)\pi_C(1 - \pi_C) + (x_C - n_C \pi_C)\pi_E(1 - \pi_E)$$

To find the score equation (equation (??)), set $\frac{\partial}{\partial \pi_C} \log L_{H_0}(\pi_C|\delta)$ equal to 0:

$$(x_E - n_E \tilde{\pi}_E)\tilde{\pi}_C(1 - \tilde{\pi}_C) + (x_C - n_C \tilde{\pi}_C)\tilde{\pi}_E(1 - \tilde{\pi}_E) \equiv 0 \quad (6)$$

To find equation (??), take the expectation of the score equation (equation (??)) under H_A :

$$E_{H_A}[(x_E - n_E \tilde{\pi}_E)\tilde{\pi}_C(1 - \tilde{\pi}_C) + (x_C - n_C \tilde{\pi}_C)\tilde{\pi}_E(1 - \tilde{\pi}_E) \equiv 0]$$

$$n_E (\pi - \vec{\pi}_E)\vec{\pi}_C(1 - \vec{\pi}_C) + n_C (\pi - \vec{\pi}_C)\vec{\pi}_E(1 - \vec{\pi}_E) \equiv 0$$

$$\pi^N = \vec{\pi}_E \left[\frac{n_E \vec{\pi}_C(1 - \vec{\pi}_C)}{n_E \vec{\pi}_C(1 - \vec{\pi}_C) + n_C \vec{\pi}_E(1 - \vec{\pi}_E)} \right] + \vec{\pi}_C \left[\frac{n_C \vec{\pi}_E(1 - \vec{\pi}_E)}{n_E \vec{\pi}_C(1 - \vec{\pi}_C) + n_C \vec{\pi}_E(1 - \vec{\pi}_E)} \right]$$

To find equation (??), invert the expectation under H_0 of the negative derivative of

$\frac{\partial}{\partial \pi_C} \log L_{H_0}(\pi_C|\delta)$ and take the square root:

$$\frac{\partial^2}{(\partial \pi_C)^2} \log L_{H_0}(\pi_C|\delta) = -\frac{x_E}{(\tilde{\pi}_E)^2} + \frac{n_E - x_E}{(1 - \tilde{\pi}_E)^2} - \frac{x_C}{(\tilde{\pi}_C)^2} + \frac{n_C - x_C}{(1 - \tilde{\pi}_C)^2}$$

$$E_{H_0} \left\{ -\frac{\partial^2}{(\partial \pi_C)^2} \log L_{H_0}(\pi_C|\delta) \equiv 0 \right\} = \frac{n_E}{\tilde{\pi}_E} - \frac{n_E}{1 - \tilde{\pi}_E} + \frac{n_C}{\tilde{\pi}_C} - \frac{n_C}{1 - \tilde{\pi}_C}$$

$$E_{H_0} \left\{ -\frac{\partial^2}{(\partial \pi_C)^2} \log L_{H_0}(\pi_C|\delta) \equiv 0 \right\}^{-0.5} = \left\{ \frac{\tilde{\pi}_E(1 - \tilde{\pi}_E)/\gamma + \tilde{\pi}_C(1 - \tilde{\pi}_C)}{n_C} \right\}^{0.5} \equiv \frac{\tilde{\sigma}_0(\hat{\delta})}{\sqrt{n_C}}$$

To find equation (??) re-express equation (??) as a function of $\vec{\pi}_C$ and to find equation (??)

re-express equation (??) as a function of π^N (equation (??)).

ACKNOWLEDGEMENTS

Financial support for the research described in this manuscript was provided by a grant from the Commonwealth Fund (20070769). The author is very appreciative of comments on manuscript drafts provided by Charles E. McCulloch.

References

1. Dibra A, Kastrati A, Mehilli J, Pache J, Schhlen H, von Beckerath N, Ulm K, Wessely R, Dirschinger J, Schmig A; ISAR-DIABETES Study Investigators. Paclitaxel-eluting or sirolimus-eluting stents to prevent restenosis in diabetic patients. *New England Journal of Medicine* 2005; **353**: 663-70. DOI: 10.1056/NEJMoa044372.
2. Bolla M, de Reijke TM, Van Tienhoven G, Van den Bergh AC, Oddens J, Poortmans PM, Gez E, Kil P, Akdas A, Soete G, Kariakine O, van der Steen-Banasik EM, Musat E, Pirart M, Mauer ME, Collette L; EORTC Radiation Oncology Group and Genito-Urinary Tract Cancer Group. Duration of androgen suppression in the treatment of prostate cancer. *New England Journal of Medicine* 2009; **360**: 2516-27. DOI: 10.1056/NEJMoa0810095.
3. Muss HB, Berry DA, Cirincione CT, Theodoulou M, Mauer AM, Kornblith AB, Partridge AH, Dressler LG, Cohen HJ, Becker HP, Kartcheske PA, Wheeler JD, Perez EA, Wolff AC, Gralow JR, Burstein HJ, Mahmood AA, Magrinat G, Parker BA, Hart RD, Grenier D, Norton L, Hudis CA, Winer EP; CALGB Investigators. Adjuvant chemotherapy in older women with early-stage breast cancer. *New England Journal of Medicine* 2009; **360**: 2055-65. DOI: 10.1056/NEJMoa0810266.
4. Sacco RL, Diener HC, Yusuf S, Cotton D, Ounpuu S, Lawton WA, Palesch Y, Martin RH, Albers GW, Bath P, Bornstein N, Chan BP, Chen ST, Cunha L, Dahlf B, De Keyser J, Donnan GA, Estol C, Gorelick P, Gu V, Hermansson K, Hilbrich L, Kaste M, Lu C, Machnig T, Pais P, Roberts R, Skvortsova V, Teal P, Toni D, Vandermaelen C, Voigt T, Weber M, Yoon BW; PRoFESS Study Group. Aspirin and extended-release dipyridamole versus clopidogrel for recurrent stroke. *New England Journal of Medicine* 2008; **359**: 1238-51. DOI: 10.1056/NEJMoa0805002.

5. Cunningham D, Starling N, Rao S, Iveson T, Nicolson M, Coxon F, Middleton G, Daniel F, Oates J, Norman AR; Upper Gastrointestinal Clinical Studies Group of the National Cancer Research Institute of the United Kingdom. Capecitabine and oxaliplatin for advanced esophagogastric cancer. *New England Journal of Medicine* 2008; **358**: 36-46. DOI: 10.1056/NEJMoa073149.
6. Home PD, Pocock SJ, Beck-Nielsen H, Gomis R, Hanefeld M, Jones NP, Komajda M, McMurray JJ; RECORD Study Group. Rosiglitazone evaluated for cardiovascular outcomes—an interim analysis. *New England Journal of Medicine* 2007; **357**: 28-38. DOI: 10.1056/NEJMoa073394.
7. Fifth Organization to Assess Strategies in Acute Ischemic Syndromes Investigators, Yusuf S, Mehta SR, Chrolavicius S, Afzal R, Pogue J, Granger CB, Budaj A, Peters RJ, Bassand JP, Wallentin L, Joyner C, Fox KA. Comparison of fondaparinux and enoxaparin in acute coronary syndromes. *New England Journal of Medicine* 2006; **354**: 1464-76. DOI: 10.1056/NEJMoa055443.
8. Twelves C, Wong A, Nowacki MP, Abt M, Burris H 3rd, Carrato A, Cassidy J, Cervantes A, Fagerberg J, Georgoulas V, Hussein F, Jodrell D, Koralewski P, Krning H, Maroun J, Marschner N, McKendrick J, Pawlicki M, Rosso R, Schller J, Seitz JF, Stabuc B, Tujakowski J, Van Hazel G, Zaluski J, Scheithauer W. Capecitabine as adjuvant treatment for stage III colon cancer. *New England Journal of Medicine* 2005; **352**: 2696-704. DOI: 10.1056/NEJMoa043116.
9. Clinical Outcomes of Surgical Therapy Study Group. A comparison of laparoscopically assisted and open colectomy for colon cancer. *New England Journal of Medicine* 2004; **350**: 2050-9. DOI: 10.1056/NEJMoa032651.
10. Cannon CP, Braunwald E, McCabe CH, Rader DJ, Rouleau JL, Belder R, Joyal SV, Hill KA, Pfeffer MA, Skene AM; Pravastatin or Atorvastatin Evaluation and Infection Therapy-Thrombolysis in Myocardial Infarction 22 Investigators. Intensive versus moderate lipid lowering with statins after acute coronary syndromes. *New England Journal of*

- Medicine* 2004; **350**: 1495-504. Erratum in: *New England Journal of Medicine* 2006; **354**: 778. DOI: 10.1056/NEJMoa040583.
11. Pfeffer MA, McMurray JJ, Velazquez EJ, Rouleau JL, Kber L, Maggioni AP, Solomon SD, Swedberg K, Van de Werf F, White H, Leimberger JD, Henis M, Edwards S, Zelenkofske S, Sellers MA, Califf RM; Valsartan in Acute Myocardial Infarction Trial Investigators. Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. *New England Journal of Medicine* 2003; **349**: 1893-906. Erratum in: *New England Journal of Medicine* 2004; **350**: 203. DOI: 10.1056/NEJMoa032292.
 12. Stone GW, Lansky AJ, Pocock SJ, Gersh BJ, Dangas G, Wong SC, Witzenbichler B, Guagliumi G, Peruga JZ, Brodie BR, Dudek D, Mckel M, Ochala A, Kellock A, Parise H, Mehran R; HORIZONS-AMI Trial Investigators. Paclitaxel-eluting stents versus bare-metal stents in acute myocardial infarction. *New England Journal of Medicine* 2009; **360**: 1946-59. DOI: 10.1056/NEJMoa0810116.
 13. van Gogh Investigators, Büller HR, Cohen AT, Davidson B, Decousus H, Gallus AS, Gent M, Pillion G, Piovela F, Prins MH, Raskob GE. Idraparinux versus standard therapy for venous thromboembolic disease. *New England Journal of Medicine* 2007; **357**: 1094-104. DOI: 10.1056/NEJMoa064247.
 14. Topol EJ, Moliterno DJ, Herrmann HC, Powers ER, Grines CL, Cohen DJ, Cohen EA, Bertrand M, Neumann FJ, Stone GW, DiBattiste PM, Demopoulos L; TARGET Investigators. Do Tirofiban and ReoPro Give Similar Efficacy Trial. Comparison of two platelet glycoprotein IIb/IIIa inhibitors, tirofiban and abciximab, for the prevention of ischemic events with percutaneous coronary revascularization. *New England Journal of Medicine* 2001; **344**: 1888-94.
 15. Shiffman ML, Suter F, Bacon BR, Nelson D, Harley H, Sol R, Shafran SD, Barange K, Lin A, Soman A, Zeuzem S; ACCELERATE Investigators. Peginterferon alfa-2a and ribavirin for 16 or 24 weeks in HCV genotype 2 or 3. *New England Journal of Medicine* 2007; **357**: 124-34. DOI: 10.1056/NEJMoa066403.

16. Victor JC, Monto AS, Surdina TY, Suleimenova SZ, Vaughan G, Nainan OV, Favorov MO, Margolis HS, Bell BP. Hepatitis A vaccine versus immune globulin for postexposure prophylaxis. *New England Journal of Medicine* 2007; **357**: 1685-94. DOI: 10.1056/NEJMoa070546.
17. Mas JL, Chatellier G, Beyssen B, Branchereau A, Moulin T, Becquemin JP, Larrue V, Livre M, Leys D, Bonneville JF, Watelet J, Pruvo JP, Albucher JF, Viguier A, Piquet P, Garnier P, Viader F, Touz E, Giroud M, Hosseini H, Pillet JC, Favrole P, Neau JP, Ducrocq X; EVA-3S Investigators. Endarterectomy versus stenting in patients with symptomatic severe carotid stenosis. *New England Journal of Medicine* 2006; **355**: 1660-71. DOI: 10.1056/NEJMoa061752.
18. Reboli AC, Rotstein C, Pappas PG, Chapman SW, Kett DH, Kumar D, Betts R, Wible M, Goldstein BP, Schranz J, Krause DS, Walsh TJ; Anidulafungin Study Group. Anidulafungin versus fluconazole for invasive candidiasis. *New England Journal of Medicine* 2007; **356**: 2472-82. DOI: 10.1056/NEJMoa066906.
19. Fowler VG Jr, Boucher HW, Corey GR, Abrutyn E, Karchmer AW, Rupp ME, Levine DP, Chambers HF, Tally FP, Vighiani GA, Cabell CH, Link AS, DeMeyer I, Filler SG, Zervos M, Cook P, Parsonnet J, Bernstein JM, Price CS, Forrest GN, Ftkenheuer G, Gareca M, Rehm SJ, Brodt HR, Tice A, Cosgrove SE; S. aureus Endocarditis and Bacteremia Study Group. Daptomycin versus standard therapy for bacteremia and endocarditis caused by *Staphylococcus aureus*. *New England Journal of Medicine* 2006; **355**: 653-65. DOI: 10.1056/NEJMoa053783.
20. Petri M, Kim MY, Kalunian KC, Grossman J, Hahn BH, Sammaritano LR, Lockshin M, Merrill JT, Belmont HM, Askanase AD, McCune WJ, Heath-Holmes M, Dooley MA, Von Feldt J, Friedman A, Tan M, Davis J, Cronin M, Diamond B, Mackay M, Sigler L, Fillius M, Rupel A, Licciardi F, Buyon JP; OC-SELENA Trial. Combined oral contraceptives in women with systemic lupus erythematosus. *New England Journal of Medicine* 2005; **353**: 2550-8. DOI: 10.1056/NEJMoa051135.
21. Van Gelder IC, Hagens VE, Bosker HA, Kingma JH, Kamp O, Kingma T, Said SA,

- Darmanata JI, Timmermans AJ, Tijssen JG, Crijs HJ; Rate Control versus Electrical Cardioversion for Persistent Atrial Fibrillation Study Group. A comparison of rate control and rhythm control in patients with recurrent persistent atrial fibrillation. *New England Journal of Medicine* 2002; **347**: 1834-40. DOI: 10.1056/NEJMoa021375.
22. Gallant JE, DeJesus E, Arribas JR, Pozniak AL, Gazzard B, Campo RE, Lu B, McColl D, Chuck S, Enejosa J, Toole JJ, Cheng AK; Study 934 Group. Tenofovir DF, emtricitabine, and efavirenz vs. zidovudine, lamivudine, and efavirenz for HIV. *New England Journal of Medicine* 2006; **354**: 251-60. DOI: 10.1056/NEJMoa051871.
23. Ginzler EM, Dooley MA, Aranow C, Kim MY, Buyon J, Merrill JT, Petri M, Gilkeson GS, Wallace DJ, Weisman MH, Appel GB. Mycophenolate mofetil or intravenous cyclophosphamide for lupus nephritis. *New England Journal of Medicine* 2005; **353**: 2219-28. DOI: 10.1056/NEJMoa043731.
24. Vincenti F, Larsen C, Durrbach A, Wekerle T, Nashan B, Blancho G, Lang P, Grinyo J, Halloran PF, Solez K, Hagerty D, Levy E, Zhou W, Natarajan K, Charpentier B; Belatacept Study Group. Costimulation blockade with belatacept in renal transplantation. *New England Journal of Medicine* 2005; **353**: 770-81. DOI: 10.1056/NEJMoa050085.
25. Walsh TJ, Tepler H, Donowitz GR, Maertens JA, Baden LR, Dmoszynska A, Cornely OA, Bourque MR, Lupinacci RJ, Sable CA, dePauw BE. Caspofungin versus liposomal amphotericin B for empirical antifungal therapy in patients with persistent fever and neutropenia. *New England Journal of Medicine* 2004; **351**: 1391-402. DOI: 10.1056/NEJMoa040446.
26. Abdulla S, Oberholzer R, Juma O, Kubhoja S, Machera F, Membi C, Omari S, Urassa A, Mshinda H, Jumanne A, Salim N, Shomari M, Aebi T, Schellenberg DM, Carter T, Villafana T, Demoiti MA, Dubois MC, Leach A, Lievens M, Vekemans J, Cohen J, Ballou WR, Tanner M. Safety and immunogenicity of RTS,S/AS02D malaria vaccine in infants. *New England Journal of Medicine* 2008; **359**: 2533-44. DOI: 10.1056/NEJMoa0807773.
27. Sundar S, Jha TK, Thakur CP, Sinha PK, Bhattacharya SK. Injectable paromomycin for

- Visceral leishmaniasis in India. *New England Journal of Medicine* 2007; **356**: 2571-81. DOI: 10.1056/NEJMoa066536.
28. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, Sthle E, Feldman TE, van den Brand M, Bass EJ, Van Dyck N, Leadley K, Dawkins KD, Mohr FW; SYNTAX Investigators. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. *New England Journal of Medicine* 2009; **360**: 961-72. DOI: 10.1056/NEJMoa0804626
29. Lai CL, Gane E, Liaw YF, Hsu CW, Thongsawat S, Wang Y, Chen Y, Heathcote EJ, Rasenack J, Bzowej N, Naoumov NV, Di Bisceglie AM, Zeuzem S, Moon YM, Goodman Z, Chao G, Constance BF, Brown NA; Globe Study Group. Telbivudine versus lamivudine in patients with chronic hepatitis B. *New England Journal of Medicine* 2007; **357**: 2576-88. DOI: 10.1056/NEJMoa066422.
30. Lacroix J, Hbert PC, Hutchison JS, Hume HA, Tucci M, Ducruet T, Gauvin F, Collet JP, Toledano BJ, Robillard P, Joffe A, Biarent D, Meert K, Peters MJ; TRIPICU Investigators; Canadian Critical Care Trials Group; Pediatric Acute Lung Injury and Sepsis Investigators Network. Transfusion strategies for patients in pediatric intensive care units. *New England Journal of Medicine* 2007; **356**: 1609-19. DOI: 10.1056/NEJMoa066240.
31. Yadav JS, Wholey MH, Kuntz RE, Fayad P, Katzen BT, Mishkel GJ, Bajwa TK, Whitlow P, Strickman NE, Jaff MR, Popma JJ, Snead DB, Cutlip DE, Firth BG, Ouriel K; Stenting and Angioplasty with Protection in Patients at High Risk for Endarterectomy Investigators. Protected carotid-artery stenting versus endarterectomy in high-risk patients. *New England Journal of Medicine* 2004; **351**: 1493-501. DOI: 10.1056/NEJMoa040127.
32. Büller HR, Davidson BL, Decousus H, Gallus A, Gent M, Piovella F, Prins MH, Raskob G, van den Berg-Segers AE, Cariou R, Leeuwenkamp O, Lensing AW; Matisse Investigators. Subcutaneous fondaparinux versus intravenous unfractionated heparin in the initial treatment of pulmonary embolism. *New England Journal of Medicine* 2003; **349**: 1695-702. Erratum in: *New England Journal of Medicine* 2004; **350**: 423. DOI: 10.1056/NEJMoa035451.

33. De Boo TM, Zielhuis GA. Minimization of sample size when comparing two small probabilities in a non-inferiority safety trial. *Statistics in Medicine* 2004; **23**: 1683-1699. DOI: 10.1002/sim.1760.
34. Farrington, C.P. and Manning, G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* 1990; **9**: 1447-1454.
35. Jewell NP. *Statistics for Epidemiology*. Chapman & Hall/CRC: Boca Raton, 2004; 77-78.
36. Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985; **4**: 213-226.
37. Chow SC, Shao J, Wang H (editors). *Sample size calculations in clinical research*. Marcel Dekker: New York, 2003.
38. Pocock SJ. The pros and cons of noninferiority trials. *Fundamental and Clinical Pharmacology* 2003; **17**: 483-490.
39. Roebuck P, Kühn A. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine* 1995; **14**: 1583-1594.
40. Fleiss JL. *Statistical Methods for Rates and Proportions*. Wiley: New York, 1981.
41. Hilton JF. Designs of superiority and noninferiority trials for binary responses are noninterchangeable. *Biometrical Journal* 2006; **48**: 934-947. DOI: 10.1002/bimj.200510288.
42. Chan ISF. Power and sample size determination for noninferiority trials using an exact method. *Journal of Biopharmaceutical Statistics* 2002; **12**: 457-469.
43. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**: 169-186.
44. Hung HM, Wang SJ, O'Neill R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal* 2005; **47**: 28-36.

45. Chow SC, Shao J. On non-inferiority margin and statistical tests in active control trials.
Statistics in Medicine 2006; **25**: 1101-1113.

Table 1. Designs based on $\theta = \delta$, at $\{\alpha, 1 - \beta\} = \{.025, .80\}$. For each design, (a) provides N_δ in upper rows and $(\gamma_\delta; \bar{\pi}_C)$ in lower rows, and (b) provides $\psi(\bar{\pi}_C)$ in upper rows and (power of the analysis based on $\theta = \log \psi$, given a design based on $\theta = \delta; E[y_C]$) in lower rows.

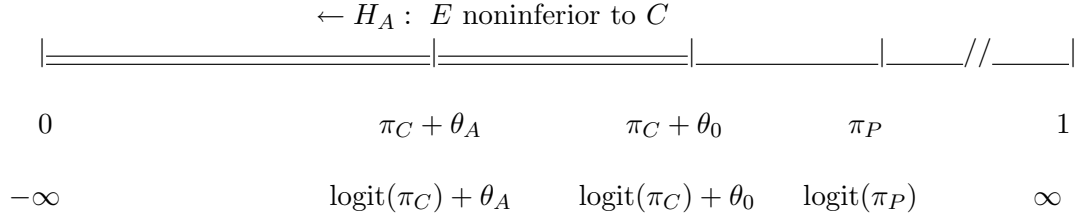
(a)	π^N	δ			
		.01	.05	.10	.20
	.01	3,266	217	80	29
		(1.95; .0058)	(3.52; .0025)	(3.99; .0020)	(3.50; .0020)
	.05	14,936	619	170	51
		(1.16; .0451)	(1.83; .0292)	(2.66; .0184)	(3.08; .0126)
	.10	28,264	1,136	288	75
		(1.05; .0951)	(1.35; .0768)	(1.74; .0582)	(2.50; .0355)
	.20	50,232	2,007	500	123
		(1.04; .1950)	(1.10; .1762)	(1.28; .1532)	(1.67; .1125)
	.40	75,344	3,008	747	182
		(1.01; .3950)	(1.01; .3754)	(1.08; .3502)	(1.16; .3012)
(b)	π^N	δ			
		.01	.05	.10	.20
	.01	2.752	22.11	56.68	129.5
		(.672; 6)	(.056; 0)	(.007; 0)	(.002; 0)
	.05	1.237	2.860	7.148	21.08
		(.795; 312)	(.683; 6)	(.382; 1)	(.093; 0)
	.10	1.118	1.746	3.041	8.369
		(.799; 1,311)	(.774; 37)	(.691; 6)	(.384; 1)
	.20	1.064	1.367	1.874	3.586
		(.800; 4,806)	(.795; 169)	(.778; 34)	(.699; 5)
	.40	1.042	1.232	1.519	2.331
		(.800; 14,818)	(.799; 565)	(.795; 126)	(.781; 25)

Table 2. Designs based on $\theta = \psi$, for the $\pi^N \times \psi$ combinations of Table 1(b). For each design, (a) provides N_ψ in upper rows and $(\gamma_\psi; \bar{\pi}_C)$ in lower rows and (b) provides $\delta(\bar{\pi}_C)$ in upper rows and (power of the analysis based on $\theta = \delta$, given a design based on $\theta = \log \psi; E[y_C]$) in lower rows.

(a)	π^N	δ			
		.01	.05	.10	.20
	.01	3,265 (.50; .0064)	600 (.27; .0019)	–	–
	.05	14,909 (.88; .0453)	632 (.53; .0315)	206 (.30; .0226)	115 (.25; .0118)
	.10	28,283 (.92; .0952)	1,141 (.74; .0784)	299 (.54; .0627)	94 (.33; .0429)
	.20	50,229 (.96; .1951)	2,019 (.87; .1767)	506 (.77; .1559)	129 (.59; .1217)
	.40	75,350 (1.00; .3950)	3,015 (.95; .3756)	753 (.95; .3513)	188 (.89; .3055)
(b)	π^N	δ			
		.01	.05	.10	.20
	.01	.0109 (.752; 14)	.0385 (.653; 1)	.0426 (.808; 0)	.0371 (.973; 0)
	.05	.0100 (.798; 359)	.0536 (.761; 13)	.1195 (.683; 4)	.1893 (.687; 1)
	.10	.0100 (.800; 1,401)	.0509 (.792; 51)	.1063 (.771; 12)	.2299 (.727; 3)
	.20	.0100 (.800; 5,000)	.0501 (.799; 191)	.1012 (.797; 45)	.2102 (.792; 10)
	.40	.0100 (.800; 14,882)	.0500 (.800; 580)	.1001 (.802; 136)	.2008 (.810; 31)

Figure 1: (a) Group-specific failure rates and log odds of failure, and (b) centered differences in relation to null (single line) and alternative (double line) noninferiority hypotheses. On the centered-difference scale, values of $\theta - \theta_0 \leq 0$ are consistent with the experimental (E) treatment being noninferior to the control (C) treatment, and values of $\theta - \theta_0 \leq \theta_A - \theta_0$ are consistent with the experimental (E) treatment being superior to the control (C) treatment. These conclusions depend on the assumption that $\pi_C < \pi_P$ although the trial does not include a placebo (P) arm.

(a) Within-groups scale: Ranges of π_E and $\text{logit}(\pi_E)$, respectively:



(b) Between-groups scale: Ranges of $\theta - \theta_0$ when $\theta = \delta$ and ψ , respectively:

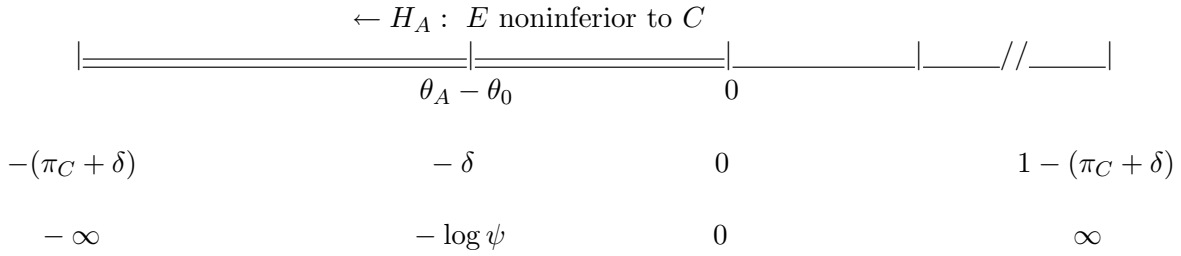
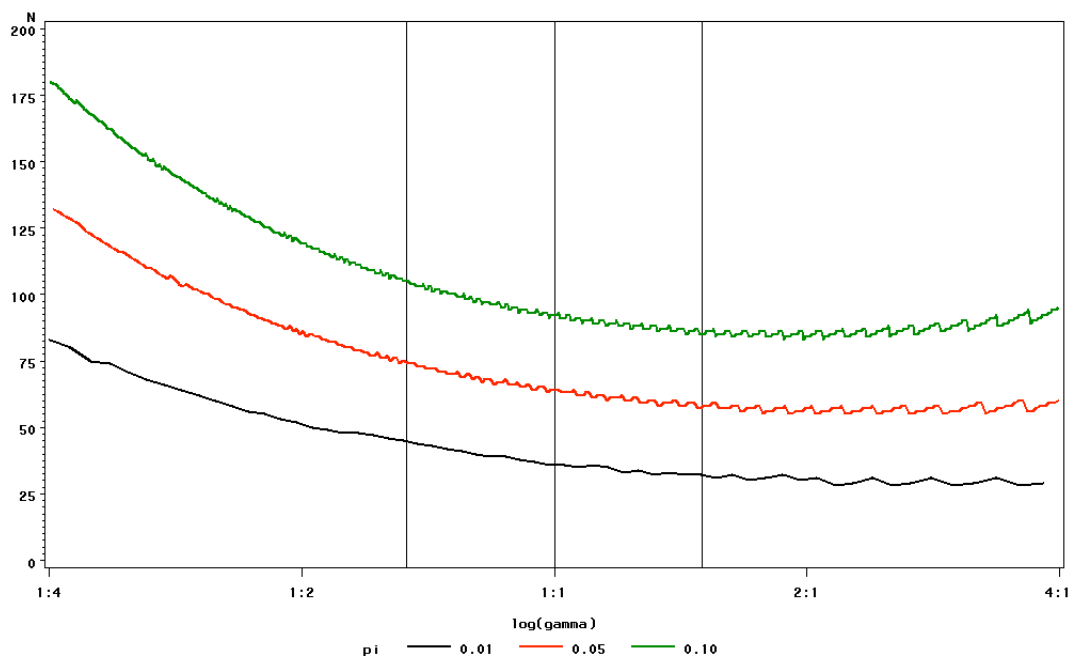
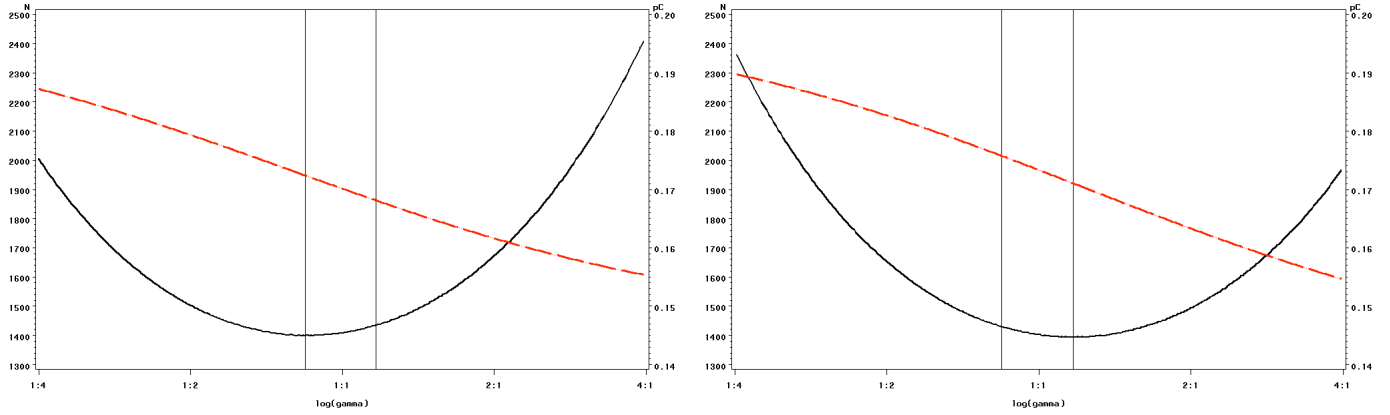


Figure 2: $N \times \log \gamma$ relationships when the noninferiority margin is parameterized on the risk-difference scale, $\theta = \delta$, with $\pi^N = .10$ (green; top), $.05$ (red; middle), and $.01$ (black; bottom), at $\{\alpha, 1 - \beta\} = \{.05, .90\}$ and $\delta = .20$. Vertical bars mark, from left, allocation ratios $\gamma = 0.67, 1.0$, and 1.5 . Instead of pre-specifying γ , our algorithm finds values of γ_δ associated with the minimum overall sample size, N_δ .

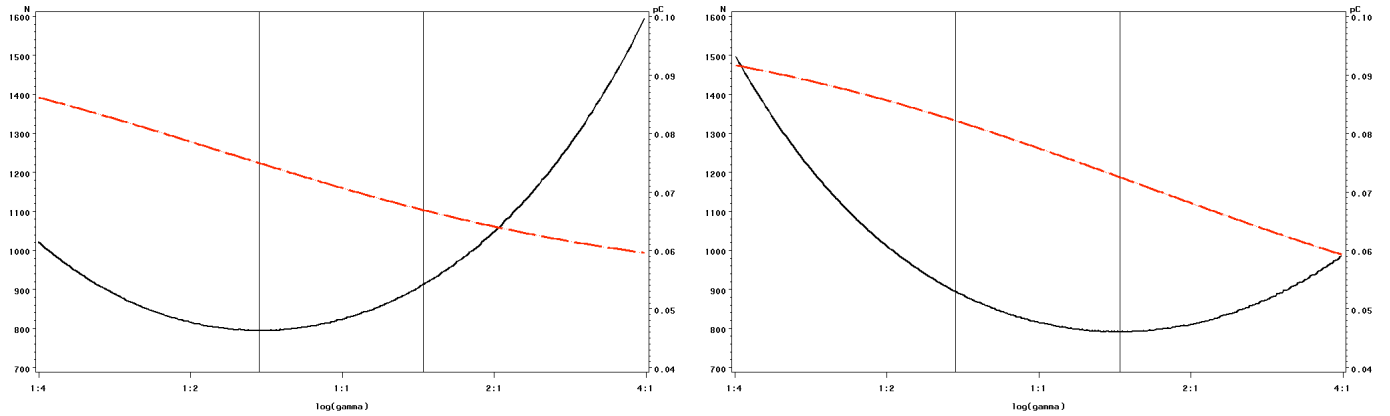


$\{\pi^N, \delta\}$	Goal	N	Range(s) of $\{\gamma, \bar{\pi}_C\}^*$
$\{.10, .20\}$	Find $\min(N)$	83	$\{1.88, .0418\} - \{2.09, .0394\}$
	Find N at $\gamma = 1.5$	85	$\{1.43, .0480\} - \{2.71, .0338\}$
	Find N at $\gamma = 1.0$	92	$(\{0.96, .0573\} - \{1.04, .0554\}), (\{3.38, .0295\} - \{3.84, .0269\})$
	Find N at $\gamma = .67$	105	$\{0.66, .0656\} - \{0.67, .0653\}$
$\{.05, .20\}$	Find $\min(N)$	55	$\{1.77, .0184\} - \{3.27, .0121\}$
	Find N at $\gamma = 1.5$	58	$\{1.33, .0218\} - \{3.84, .0107\}$
	Find N at $\gamma = 1.0$	64	$\{0.94, .0260\} - \{1.06, .0246\}$
	Find N at $\gamma = .67$	74	$\{0.65, .0304\} - \{0.68, .0299\}$
$\{.01, .20\}$	Find $\min(N)$	28	$\{2.16, .0028\} - \{3.59, .0019\}$
	Find N at $\gamma = 1.5$	32	$\{1.31, .0039\} - \{1.87, .0031\}$
	Find N at $\gamma = 1.0$	36	$\{0.99, .0046\} - \{1.11, .0043\}$
	Find N at $\gamma = .67$	45	$\{0.66, .0056\}$

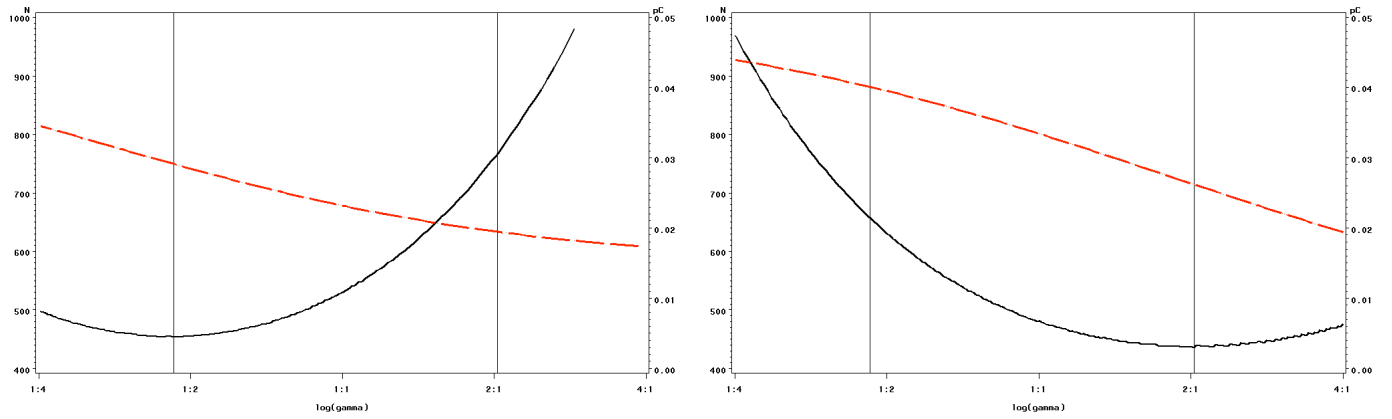
Figure 3: $N \times \log \gamma$ relationships for designs $\{\pi^N, \delta\} = \{.20, .06\}$ (top), $\{.10, .06\}$ (center), and $\{.05, .06\}$ (bottom), at $\{\alpha, 1 - \beta\} = \{.025, .80\}$. Vertical bars mark the minimum overall sample sizes when the noninferiority margin is parameterized on the log-odds scale, $\theta = \log \psi$ (left plot and bars), and when parameterized on the risk-difference scale, $\theta = \delta$ (right plot and bars). Improperly specifying $\pi^N \equiv \pi_C$ substantially underestimates the sample size.



	$\theta = \log \psi$			$\theta = \delta$		
	N	$\{\gamma, \bar{\pi}_C\}^*$	$\{\pi^N, \delta(\bar{\pi}_C)\}$	N	$\{\gamma, \bar{\pi}_C\}^*$	$\{\pi^N, \psi(\bar{\pi}_C)\}$
$\{\pi^N = .20, \psi = 1.456\}$:	1,399	(.844, .172)	$\{.200, .0603\}$	1,393	(1.17, .171)	$\{.200, 1.456\}$
$\{\pi_C = .20, \psi = 1.405\}$:						
– Midpoint Approach	1,538	(.874, .202)	$\{.230, .0604\}$	1,541	(1.13, .201)	$\{.230, 1.404\}$
– Tailored Approach	1,548	(.881, .200)	$\{.228, .0600\}$	1,536	(1.14, .200)	$\{.229, 1.406\}$



	$\theta = \log \psi$			$\theta = \delta$		
	N	$\{\gamma, \bar{\pi}_C\}^*$	$\{\pi^N, \delta(\bar{\pi}_C)\}$	N	$\{\gamma, \bar{\pi}_C\}^*$	$\{\pi^N, \psi(\bar{\pi}_C)\}$
$\{\pi^N = .10, \psi = 1.954\}$:	794	(.686, .075)	$\{.100, .0617\}$	790	(1.45, .072)	$\{.100, 1.954\}$
$\{\pi_C = .10, \psi = 1.714\}$:						
– Midpoint Approach	962	(.751, .104)	$\{.131, .0621\}$	988	(1.34, .102)	$\{.130, 1.705\}$
– Tailored Approach	995	(.752, .100)	$\{.126, .0600\}$	975	(1.32, .100)	$\{.128, 1.715\}$



	$\theta = \log \psi$			$\theta = \delta$		
	N	$\{\gamma, \bar{\pi}_C\}^*$	$\{\pi^N, \delta(\bar{\pi}_C)\}$	N	$\{\gamma, \bar{\pi}_C\}^*$	$\{\pi^N, \psi(\bar{\pi}_C)\}$
$\{\pi^N = .05, \psi = 3.506\}$:	454	(.464, .029)	{.050, .0660}	436	(2.03, .026)	{.050, 3.506}
$\{\pi_C = .05, \psi = 2.348\}$:						
– Midpoint Approach	580	(.614, .058)	{.084, .0683}	653	(1.60, .053)	{.080, 2.270}
– Tailored Approach	668	(.603, .050)	{.072, .0596}	630	(1.65, .050)	{.077, 2.341}

* Medians of solution ranges.