Published in final edited form as:

J Biopharm Stat. 2009 July ; 19(4): 685–699. doi:10.1080/10543400902964142.

Non-inferiority Trial Design and Analysis with an Ordered Three-Level Categorical Endpoint

Erica Brittain and Zonghui Hu

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda MD 20892

Abstract

This paper extends standard methodology for non-inferiority trial design from a binary endpoint to an ordered three-level endpoint: such as “success”, “intermediate”, and “failure”. A metric that summarizes outcome on this endpoint is proposed, and the corresponding sample size requirements are presented. This ordered endpoint can be collapsed into two different binary endpoints, respectively lumping “intermediate” outcomes with “success” or with “failure”. We describe how the ordered three-level endpoint compares with these two binary endpoints with respect to the non-inferiority margin and sample size requirements.

Keywords

non-inferiority; clinical trial design; ordinal data

1 Introduction

Methods for designing and analyzing non-inferiority trials with a binary endpoint are well established (Blackwelder, 1982; Makuch and Simon, 1978). For a binary endpoint, the metric often used to compare treatments is the difference in proportions, and the corresponding non-inferiority margin might be .10, say. In this paper we consider extending this approach to a three-level ordinal categorical endpoint. Comparison of two groups with a three-level endpoint in a superiority trial setting could be analyzed in a variety of ways, but in the non-inferiority trial setting, one must define a metric that can be compared statistically to some pre-specified non-inferiority margin. Furthermore, one must think about the magnitude of the chosen margin for this particular metric.

In the preliminary stages of considering designs for non-inferiority trials of MRSA (Methicillin Resistant *Staphylococcus Aureus*) skin infections, medical colleagues expressed interest in considering a three level endpoint: success, intermediate, and failure, although skin infection studies traditionally use only success and failure. They indicated that the range of outcomes of a treated skin infection is not naturally dichotomous, but rather is implicitly continuous. Furthermore, the responses naturally fall into three clusters: clear success, clear failure, and a gray zone, in between; where clear success corresponds to complete resolution of symptoms, clear failures corresponds to worsening symptoms or little or no change, and the gray zone corresponds to some important improvement in the symptoms, but clearly less than complete resolution. The medical colleagues estimated that the intermediate cluster (i.e., the gray zone) would represent a fairly large fraction of the assessments, perhaps as much as 25%. (In fact, Daly et al. 1990 presented results of a skin and skin structures infection trial in which an intermediate “improvement” category was provided; approximately 30% of the outcomes fell into this category.) Dividing the gray zone into success and failure requires the development of somewhat arbitrary criteria, and is likely not to be as reproducible across evaluators as the

trichotomous endpoint. Of course, one could lump together the success and intermediate categories to create a binary endpoint “response” and “no response”, or likewise lump intermediate with failure to create “success” and “not success”. However, the investigators felt the intermediate category was distinct, and might wish to distinguish this from the other two categories. Furthermore, the investigators felt it was reasonable to view this intermediate category as roughly half way between the two other categories. Our colleagues ultimately did not seriously consider the three-level approach, partly because the relevant statistical issues for a non-inferiority design had not been fully worked out. This motivated us to develop such an approach.

Clinical trials require pre-specification of the endpoint. If success rates were the same for two drugs, and there were differences at the intermediate level, a clinical trial that pre-specified the success endpoint might “demonstrate” non-inferiority, but this would fail to acknowledge the potentially meaningful differences between these drugs. Alternatively, two drugs might have similar response rates, but differences on success, and if the response rate were the pre-specified outcome, this too would fail to acknowledge a potentially meaningful difference. Thus, our goal is to define a metric and method that would be sensitive to either of these differences.

The method is described in Section 2, a brief example is presented in Section 3, and a discussion is provided in Section 4.

2 Proposed Approach

We define the following notation. P_{SC} is the true proportion of success in the control group, P_{IC} is likewise for the proportion with intermediate outcomes, and P_{RC} is the true proportion in the control group with response outcomes (i.e., success and intermediate outcomes combined), so that $P_{RC}=P_{SC}+P_{IC}$. Notation for the experimental arm is likewise: P_{SE} and so forth.

Non-inferiority clinical trials of anti-infectives typically use the difference in proportions as the measure of treatment effect. Thus, the true treatment difference for the binary success endpoint is

$$\theta_s = P_{SE} - P_{SC}.$$

The true treatment difference for the binary “response” endpoint is

$$\theta_r = (P_{SE} + P_{IE}) - (P_{SC} + P_{IC}).$$

In a non-inferiority trial, if one of these were selected as the primary endpoint, its corresponding point estimate and confidence interval would be compared to some selected margin, Δ .

As Agresti (2002) noted in the introduction to his classic text on ordinal data, it is often constructive to assign scores that approximate the underlying continuous scale. So, for the three-level endpoint we can assign a score to each of the three outcomes, and compare mean scores between the two groups. While any scoring system could be considered, we found the following approach the most appealing. As a natural extension of the binary endpoint, we assign the value of 0 for failure, 1 for success, and a value of p for the intermediate category, where p is a specified value between 0 and 1 that represents where the intermediate level is deemed to fall on the continuum between success and failure. When $p=0$, then this score reduces to the binary approach when the intermediate group is lumped with failure, and when $p=1$, then this score reduces to the binary approach where intermediates are lumped with success. The

expected mean score for the control group is consequently defined as $E[X_C] = P_{SC} + \rho(P_{IC})$, where P_{SC} is the true proportion with outcomes of success in the control group; $E[X_E]$ is defined likewise. Then, the expected treatment effect with the three-level score endpoint is as follows:

$$\theta_3 = \{P_{SE} + \rho(P_{IE})\} - \{P_{SC} + \rho(P_{IC})\}.$$

We note that this is equal to $(P_{SE} - P_{SC}) + \rho(P_{IE} - P_{IC})$, which can be rewritten as $(1 - \rho)(P_{SE} - P_{SC}) + \rho\{(P_{SE} + P_{IE}) - (P_{SC} + P_{IC})\}$. Thus, it is interesting to recognize that

$$\theta_3 = (1 - \rho)\theta_s + \rho\theta_r.$$

That is, the true treatment difference of interest with this three-level endpoint is a simple weighted average of the two binary treatment differences of interest, and can be viewed as a compromise.

This method requires specifying a value for ρ ; Agresti (2002) notes that assigning scores “requires good judgment and guidance from the researchers who use the scale”. In the skin infection clinical trial mentioned above, the investigators considered the intermediate outcome as roughly half-way between success and failure, thus we would assign the intermediate outcome a value of .5. We note that for $\rho = .5$, $\theta_3 = (\theta_s + \theta_r)/2$, a simple average of the two binary treatment differences. We emphasize that one cannot use the internal data to estimate ρ ; it is an inherently subjective judgment about where the intermediate outcome falls on the continuum. If it is understood to lie closer to success than failure, that is, it is close to a complete resolution and much better than a failure, then $\rho = .75$ might be a reasonable choice, and vice versa. There may be cases where external data could be employed as a basis for setting ρ . For example, imagine a setting where some proportion of intermediates were known to eventually become successes and there were an estimate of this proportion; then ρ might be set equal to this estimate. But, absent such a scenario, the decision would be based on clinical judgment about where the intermediate cluster lies on the continuum, and if there is no clear basis for assigning a score greater than .5 or a score less than .5, then equally spaced scoring (i.e., $\rho = .5$) seems like the natural choice. Thus, except where noted otherwise, we focus on the $\rho = .5$ case for the rest of the paper, as we suspect this would likely be the selected value in the majority of applications. The clear interpretation and clinical relevance of the chosen metric of the primary endpoint is critically important. We believe the mean score with a scoring system of 1 for success, .5 for intermediate, and 0 for failure is easily understood, and has natural interpretation, in those settings where it is reasonable to view the intermediate response as mid-way between the other outcomes, and the metric is likewise natural for values of ρ in general.

2.1 Confidence Interval for θ_3

Let p_{SE} and p_{IE} denote the observed proportions in a non-inferiority study in the experimental group, and likewise for the control group. Let the corresponding sample sizes be n_E and n_C . Then the estimated treatment effect of θ_3 , when $\rho = .5$, is

$$D_3 = \{p_{SE} + .5(p_{IE})\} - \{p_{SC} + .5(p_{IC})\}.$$

The variance for the estimate D_3 is $\text{Var}(p_{SE} + .5(p_{IE})) + \text{Var}(p_{SC} + .5(p_{IC}))$. The first term is equal to $\text{Var}(p_{SE}) + \text{Var}(p_{IE})/4 + \text{Cov}(p_{SE}, p_{IE})$. Under the multinomial distribution, the covariance of p_{SE} and p_{IE} is $-(P_{SE})(P_{IE})$. Thus,

$$\text{Var}(D_3) = [P_{SE}(1 - P_{SE}) + \{P_{IE}(1 - P_{IE})/4\} - P_{SE}P_{IE}]/n_E + [P_{SC}(1 - P_{SC}) + \{P_{IC}(1 - P_{IC})/4\} - P_{SC}P_{IC}]/n_C.$$

Thus, using the normal distribution approximation, we can estimate the bounds of the 95% confidence interval for θ_3 as $D_3 \pm 1.96 \sqrt{\text{Var}(D_3)}$. Note that all confidence intervals in this paper use the normal approximation, and thus the results apply only to those settings where the normal distribution will yield a reasonable approximation, that is, where the sample size is sufficiently large, preferably at least 100 per group, and the quantities to be estimated, $p_{SE} + .5(p_{IE})$ and $p_{SC} + .5(p_{IC})$, are not very close to zero or one. We further note that for general ρ , $D_3 = \{p_{SE} + \rho(p_{IE})\} - \{p_{SC} + \rho(p_{IC})\}$, and the formula for variance generalizes to:

$$\text{Var}(D_3) = [P_{SE}(1 - P_{SE}) + \rho^2 \{P_{IE}(1 - P_{IE})\} - 2\rho P_{SE}P_{IE}]/n_E + [P_{SC}(1 - P_{SC}) + \rho^2 \{P_{IC}(1 - P_{IC})\} - 2\rho P_{SC}P_{IC}]/n_C.$$

Thus, non-inferiority is considered demonstrated if the confidence interval lies completely above $-\Delta_3$, where Δ_3 is the pre-specified margin. We note that this is exactly equivalent to whether or not the corresponding hypothesis test statistic, $Z = (D_3 + \Delta_3) / \sqrt{\text{Var}(D_3)}$, exceeds 1.96; that is, if it is statistically significant. The null hypothesis is $\theta_3 < -\Delta_3$ and the alternative hypothesis is $\theta_3 > -\Delta_3$. Blackwelder (2005) recommends consideration of the confidence interval results over the hypothesis test results, as the former is focused on estimation, and informs about which values of θ_3 “are consistent with the data” as opposed to focusing on the strength of the evidence associated with a particular margin. Since non-inferiority trials in infectious disease traditionally employ the confidence interval framework, this paper takes that approach as well.

2.2 Selection of Non-inferiority Margin for θ_3

Every non-inferiority trial requires pre-specification of a margin. If the confidence interval for the specified treatment difference of interest excludes this margin, then the trial has demonstrated non-inferiority. For example, say the confidence interval for θ_3 , (i.e., the mean score in the experimental group minus the mean score in the control group), is $(-.07, .12)$. If the non-inferiority margin was specified as $.10$, then non-inferiority has been demonstrated, because the confidence interval lay completely above $-.10$. In other words, the trial has ruled out any possible inferiority of the experimental treatment to the control with respect to this endpoint that is greater than $.10$.

Selection of the non-inferiority margin has been discussed by a number of authors (Weins, 2002; D'Agostino et al., 2003; Hung et al., 2003). The choice of the non-inferiority margin includes a subjective element, specifically, a determination, based on medical judgment, must be made about the largest clinically acceptable loss in efficacy with respect to the endpoint of interest. For example, if the endpoint were mortality, presumably only the slightest decline in efficacy might be acceptable, and then, perhaps, only if there were some other benefits of the new drug. However, if the endpoint related to a bad outcome that could be easily and completely reversed by further treatment, then a larger margin could be considered.

In the regulatory environment, the margin must also be smaller than the known magnitude of the treatment effect of the active control drug over placebo, *with respect to the same endpoint*. So effectively, the margin is the smaller of the largest acceptable loss and the treatment effect of the active control. This minimizes the possibility that a drug that is no better

than a placebo can be licensed on the basis of being “non-inferior” to a drug that is known to have efficacy. Some have written about specifying the margin to preserve or retain a certain fraction of the benefit, such as half of the benefit (Hung et al., 2003; ICH E-10, 2000).

Now consider the case where we assume the active control treatment has a very large treatment effect, so that the choice of the margin is essentially only a function of the largest clinically acceptable loss. Suppose an investigative team decided that .10 was a sufficient standard for the non-inferiority margin in the binary endpoints, but now wants to consider using a three-level endpoint. What would be a corresponding measure of similarity for the three-level endpoint? To address this, it is useful to examine explicitly the non-inferiority regions for a given endpoint and margin. For the $\rho=.5$ case, we can define such regions on the $P_{SE}-P_{IE}$ plane for a set of specified true values of P_{SC} and P_{IC} as follows, where Δ_S is the non-inferiority margin chosen for the success binary endpoint, and likewise for the other endpoints:

$$\begin{aligned}\eta_S &= \{(P_{SE}, P_{IE}) : P_{SE} > P_{SC} - \Delta_S\}, \\ \eta_R &= \{(P_{SE}, P_{IE}) : P_{SE} + P_{IE} > P_{SC} + P_{IC} - \Delta_R\}, \text{ and} \\ \eta_3 &= \{(P_{SE}, P_{IE}) : P_{SE} + .5P_{IE} > P_{SC} + .5P_{IC} - \Delta_3\}.\end{aligned}$$

These regions of non-inferiority are plotted for two examples in Figure 1, where the non-inferiority margin is .10 for all three endpoints. Not surprisingly, in the $\rho=.5$ case, the non-inferiority region of the three-level endpoint seems to be a compromise between that of the other two endpoints, so that if an investigator is comfortable with .10 for each of the binary endpoints, it is probably an appropriate margin for the three-level endpoint. That said; if the true response rate and the true success rate for the active control drug are very different, .99 and .50, say, then it may not be reasonable to select the same margin for these two binary endpoints. We also see in the example shown in Figure 1A, the zone of clinical non-inferiority for the three-level endpoint is considerably more generous than the success endpoint, and the response endpoint is even more generous, when the same margin is used. However, in Figure 1B, which considers a different set of P_{SC} and P_{IC} , there is more of an even trade-off between the endpoints. Other examples exist where the success endpoint is the most generous. Thus, there are no simple answers; investigators need to think carefully about the appropriate clinical indifference zone for any given endpoint and situation.

2.3 Comparison of Sample Size Requirements of the Three-level Endpoint with Binary Endpoints

The comparison of necessary sample size would depend on the choice of the margin and specific parameters of the two populations being compared. For the purpose of sample size estimation, let us assume that the true outcome probabilities are the same in both groups, and let P_S denote the true probability of success in either group, P_I denote the true probability of intermediate response in either group, and let P_R be likewise. Thus the sample size per group under equal allocation for each of the endpoints is as follows, where Z_β is the standard normal deviate associated with the $(1-\alpha)$ confidence interval and Z_β is the standard normal deviate associated with Power $(1-\beta)$:

$$\begin{aligned}n_S &= \left\{ \frac{(Z_\alpha + Z_\beta)}{\Delta_S} \right\}^2 \{2P_S(1 - P_S)\}, \\ n_R &= \left\{ \frac{(Z_\alpha + Z_\beta)}{\Delta_R} \right\}^2 [2\{(P_S + P_I)(1 - (P_S + P_I))\}], \\ n_3 &= \left\{ \frac{(Z_\alpha + Z_\beta)}{\Delta_3} \right\}^2 [2\{P_S(1 - P_S) + P_I(1 - P_I) / 4 - P_S P_I\}] \text{ for } \rho=.5, \text{ and} \\ n_3 &= \left\{ \frac{(Z_\alpha + Z_\beta)}{\Delta_3} \right\}^2 [2\{P_S(1 - P_S) + \rho^2 P_I(1 - P_I) - 2\rho P_S P_I\}] \text{ in general.}\end{aligned}$$

Obviously, as these equations indicate, the sample size requirements are strongly influenced by the choice of the margin. We further note that if the two treatments are not assumed equal, this can impact the sample size result substantially. For example, for general θ , and different treatment parameters, the per group sample size formula for the three-level endpoint is:

$$n_3 = \left\{ \left(Z_{\alpha} + Z_{\beta} \right) / \left(P_{SC} + \rho P_{IC} - P_{SE} - \rho P_{IE} + \Delta_3 \right) \right\}^2 \left[\left\{ P_{SE} (1 - P_{SE}) + \rho^2 P_{IE} (1 - P_{IE}) - 2\rho P_{SE} P_{IE} \right\} + \left\{ P_{SC} (1 - P_{SC}) + \rho^2 P_{IC} (1 - P_{IC}) - 2\rho P_{SC} P_{IC} \right\} \right].$$

2.3.1 Relative Sample Size Requirements to Preserve a Specific Proportion of the Benefit of the Treatment—We first consider the case where we allow the margin to be different for the three endpoints, and where the margin is designed to preserve a specified percentage of the active control treatment benefit over placebo; these values are determined from historic data. For example, a non-inferiority study might be designed to show the new drug retains at least 50% of the treatment benefit of the active control over placebo, in terms of the endpoint to be used. For simplicity, we focus on the idealized scenario where the magnitude of the benefit is exactly known. That is, we know $(P_S - P_S^*)$, where P_S^* is the true proportion of success with placebo, and we know $(P_R - P_R^*)$, where P_R^* is the true proportion of response with placebo. Let λ denote the proportion of benefit to be preserved, such as .50, in the example above. Then, Δ_S is specified to be $\lambda(P_S - P_S^*)$, Δ_R is $\lambda(P_R - P_R^*)$, and Δ_3 is $\lambda \{ (P_S - P_S^*) + \rho(P_R - P_R^*) \}$. The relative sample sizes based on the equations in Section 2.3 are a function of the ratio of the variances and the ratio of the margins. Using algebra, we determine that the relative sample size is a function of the ratio of the variances and γ , where

$\gamma = \left(\frac{P_R - P_R^*}{P_S - P_S^*} \right)$, because the ratio of the non-inferiority margins are all simple functions of γ . We note that γ is one when the active control's treatment benefit is equal for the two binary endpoints, and it is greater than one when the treatment benefit between active control and placebo is greater for the response endpoint than the success endpoint, and vice versa. The relative sample size requirements are independent of the chosen preserved benefit, λ . They are:

$$\begin{aligned} \frac{n_S}{n_3} &= \left\{ \frac{P_S(1-P_S)}{P_S(1-P_S) + \rho^2 P_I(1-P_I) - 2\rho P_S P_I} \right\} (1 - \rho + \rho\gamma)^2, \\ \frac{n_R}{n_3} &= \left\{ \frac{(P_S + P_I)(1-P_S - P_I)}{P_S(1-P_S) + \rho^2 P_I(1-P_I) - 2\rho P_S P_I} \right\} \left(\frac{1-\rho}{\gamma} + \rho \right)^2, \text{ and} \\ \frac{n_S}{n_R} &= \left\{ \frac{P_S(1-P_S)}{(P_S + P_I)(1-P_S - P_I)} \right\} \gamma^2. \end{aligned}$$

We can define the region which is associated with the endpoint with the smallest sample size: $\Omega_S = \{ (P_S, P_I) : n_S < \min(n_R, n_3) \}$, and Ω_R and Ω_3 are defined likewise. Figure 2 presents these regions for three different values of γ , for the case of equal preservation of benefit, for the $\rho = .5$ case. These plots illustrate that each of the three endpoints have substantial regions where it is associated with the smallest sample size.

2.3.2 Relative Sample Size Requirements for Equal Non-inferiority Margins—

Now let us consider the case where the design team has decided to employ the same margin, Δ , regardless of which of the three endpoints is used. This might be the case where the treatment effect over placebo is large, and the goal of the selection of the non-inferiority margin is to specify the largest difference that is clinically acceptable. If the specified margin for all of the endpoints would be equal, then the sample size comparison is simply a function of the predicted variance. The second plot in Figure 2, which corresponds to the $\gamma=1$ case, also provides the regions of the smallest sample size requirements, for the case where the same margin would

be used for each of the three endpoints and where $\rho=.5$. (We emphasize that these relationships apply whenever the margin is chosen to be equal for all endpoints, regardless of the actual value of γ .) Relative sample size for equal margins (as illustrated by the second plot in Figure 2 for the $\rho=.5$ case) corresponds to the following equations:

P_S and P_I Relationship	Smallest Sample Size
$P_I < 1 - (2/\rho)P_S$	n_S
$1 - (2/\rho)P_S < P_I < 1 - \{2/(1+\rho)\}P_S$	n_3
$P_I > 1 - \{2/(1+\rho)\}P_S$	n_R

The relative sample size depend on the values of P_I and P_S . Simply put, for the $\rho=.5$ case, when $P_S > .75$, then the response endpoint will lead to the smallest sample size, whereas only when $P_S < .25$, can the success endpoint result in the smallest sample size, yet the three-level endpoint often produces the smallest sample size where $P_S < .75$. Interestingly, there is no case where n_3 is the largest of the trio of sample sizes, but both of the binary endpoints can be the largest. Table 1 provides some relative sample size requirements for the equal margin case as a function of the true success and intermediate probabilities.

2.4 Simulation Results

In order to confirm that the sample size calculation, power and Type I error are as expected, we conducted a simulation for a range of parameters; results are shown in Table 2. We conducted 10,000 replications when estimating power and 100,000 replications when estimating rejection rates under the null. Estimated rejection rates were statistically consistent with the nominal levels for simulated parameters, with several exceptions. In several cases (5 and 6), the simulated power was slightly significantly greater than .90; this might be related to the fact that the sample size is always rounded up. In addition, the simulated rejection rate for Case 9 was .0271, which is close, but statistically larger than .025. It is noted that this corresponds to the smallest sample size considered (153 per group), and is consistent with the simple binary case, where elevated Type I error is expected for smaller sample sizes. All analogous simulations were also performed for $\rho=.3$ and .7, with very similar results (not shown). In conclusion, this simulation study suggests that the formulas provided in the paper yield reasonable statistical operating procedures. However, there is some potential for anti-conservative results for relatively small sample sizes. We speculate that once sample size is as great as 200 per group, this concern diminishes.

2.5 Alternate Approach to Estimation of Variance

Some researchers recommend that the approach of Farrington and Manning (1990) be used in the typical binary non-inferiority setting. This method uses a maximum likelihood estimate for the variance constrained under the null, and Roebruck and Kuhn (1995) found that this method often had better operating characteristics than the simple approach that uses the within group estimates of proportions in the variance formulas. In the typical binary non-inferiority case, Farrington and Manning (1990) derived closed form solutions for this alternate variance formulation, to allow straightforward use of this method. One might want to modify the three-level procedure similarly. However, the greater complexity does not appear to allow a closed-form solution. Implementation involves the following steps. Let n_{SE} denote the number of successes observed in the experimental group and likewise, then estimates under the null for P_{SE} , P_{IE} , and P_{IC} are found by maximizing the following likelihood, the estimate for P_{SC} is a simple function of these estimates, and the maximization is restricted such that all sole and combined estimated proportions fall between 0 and 1. The likelihood is:

$$P_{SE}^{n_{SE}} P_{IE}^{n_{IE}} (1 - P_{SE} - P_{IE})^{n_{FE}} (P_{SE} + \rho P_{IE} - \rho P_{IC} - \theta_3)^{n_{SC}} P_{IC}^{n_{IC}} (1 - P_{IC} - P_{SE} - \rho P_{IE} + \rho P_{IC} + \theta_3)^{n_{FC}}.$$

The variance formula in Section 2.1 would then use these estimates instead of the usual observed proportions.

Attempts to use commercial software to determine this variance estimate were not always successful because of failures to converge, or complex restricted bounds. We were able to find numerical solutions using the same parameters and sample size as simulations presented in Section 2.4; however, for most of these parameters, very little difference was seen in variance estimation, and hence in rejection rates. For example, with Case 1, with $n=225$, the mean variance of D_3 was .000946 with the simple formula and .000954 with the Farrington-Manning (FM) like variance formula, and differences in power were negligible: .900 versus .901. Similarly, with the corresponding null example (Case 7), the mean variances were .001056 and .001059 respectively, with rejection probabilities of .0249 and .0248, respectively. The one exception was Case 9 where the Type I error was estimated to be .0271 with the simple formula, which was statistically larger than .025, but it was .0256 using a FM like approach. Not surprisingly, this case had the smallest sample size ($n=153$) considered. When replications that yield discordant rejections were examined, the test statistic which corresponds exactly to the confidence interval approach (i.e., $(D_3 + \Delta_3) / \sqrt{\text{Var}(D_3)}$) was generally just slightly greater than 1.96 for the simple formula and slightly less than 1.96 for the FM like approach. Thus, when using the simple formula with a modest sample size case, borderline demonstrations of non-inferiority should be interpreted cautiously.

We also considered a non-MLE alternate version, analogous to the second method proposed in the original binary case paper by Farrington and Manning (1990); which allowed a straightforward comparison to the simple approach. Using this strategy, $\tilde{P}_{SE} = [P_{SE} + P_{SC} + \Delta/(1+\rho)]/2$; $\tilde{P}_{IE} = [P_{IE} + P_{IC} + \Delta/(1+\rho)]/2$; $\tilde{P}_{SC} = [P_{SE} + P_{SC} - \Delta/(1+\rho)]/2$; $\tilde{P}_{IC} = [P_{IE} + P_{IC} - \Delta/(1+\rho)]/2$, these lead to estimates that are consistent with the null, however these are not ML estimates. Algebraically it can be shown that $V_S - V_{FM*} = (\Delta^2 - D_3^2)/(2n)$, where n is the per group sample size, and V_S is the estimate of $\text{Var}(D_3)$ using our original paper's unconstrained simple estimates and V_{FM*} is the variance using this alternate non-MLE FM like approach. Given that $E[D_3^2] = \text{Var}(D_3) + \theta_3^2$, one can determine, under the null, that $E[V_{FM*}]/E[V_S]$ equals approximately $(1 + 1/(2n))$. Even though this relationship is based on the alternate non-MLE FM like method, it appeared very consistent with our null case simulation results based on the MLE FM like approach. For example, in Case 9, the ratio of mean variances was 1.003268 which equals $[1 + 1/\{(2)(153)\}]$. Thus, this finding further supports the conclusion that the difference between constrained and unconstrained variances even for moderate sample sizes is small.

Thus, in most moderate to large sample size situations, we suspect the extra effort involved in employing the Farrington-Manning like approach would not be worthwhile. However, in the relatively less common scenario of a small sample size case, the Farrington Manning like approach could be considered, as it would probably provide better control of Type I error.

3 Example

Published data from a clinical trial allow us to consider the three different endpoints with a real world example. We considered results from a published study that compared once-daily gatifloxacin (which we denote “Experimental”) versus three-times-daily co-amoxiclav (which we denote “Control”) in the treatment of community acquired pneumonia (Lode et al., 2004). The primary outcome of clinical response was assessed by three ordinal categories: “cure”,

“improvement”, and “non-response”; the trial’s specified primary endpoint was binary response (i.e., cure plus improvement). While the definitions of the three outcome classifications were more complex, cure was associated with “complete resolution of acute signs and symptoms of pneumonia”, improvement with “resolution of >50% of all signs and symptoms”, and failure with “lack of resolution of >50% of all signs and symptoms”. Treatment lasted 5-10 days in both arms, and the primary assessment time was at the end of treatment; the end of study assessment was two to four weeks later. The results for each of the three possible endpoints are presented in Table 3, where ρ is set to .5. In this example, all endpoints led to very similar results at the end of treatment which was the primary assessment time; however, the three-level endpoint did provide the greatest evidence of non-inferiority. This trial used a margin of .15, but had an extremely stringent margin of .01 been chosen, only the three-level endpoint would have met this standard. The three-level endpoint had the greatest lower bound, because it combined a relatively small variance with a relatively large treatment effect point estimate. At the secondary assessment time (i.e., end of study), the pre-specified endpoint, binary response, provided the greatest evidence of non-inferiority. (Note: the confidence intervals for response presented in the published paper differ slightly from those presented in Table 3 since a different approach to confidence interval calculation was used.)

At both assessment times, the width of the confidence intervals for the binary response and three-level-endpoint were very similar, whereas the width associated with binary success was clearly larger. This is consistent with the middle plot in Figure 2, since, under equivalence, the data suggest P_S is .52 and P_I is .34 at the end of treatment assessment and .65 and .18 at the end of study assessment. Both of these points lie very close to the border between Ω_R and Ω_3 , that is, the border between the area where the response endpoint minimizes sample size and the area where the three-level endpoint leads to smallest sample size.

Unlike the end of treatment results, the end of study results for the binary success and binary response endpoints are quite different, with the response endpoint suggesting a difference that is two to three times that of the success endpoint. In a setting where, *a priori*, it is unclear which of these two binary endpoints is more likely to reveal meaningful differences between treatments, the three-level endpoint is a relatively low risk alternative that will represent an average of the two endpoints. But, of course, the choice should depend on what is clinically meaningful: if investigators believe that cure and improvement proportions are both important, but should be distinguished, then the three-level endpoint should be considered.

4 Discussion

We have determined regions where each method has the smallest sample size, under a variety of circumstances. That said; we believe the decisive factor will often be the clinical relevance of the endpoint, not which endpoint is associated with the smallest sample size. If the intermediate category is halfway between the worst and best outcomes, and cannot be lumped with either of the other outcomes without losing important information, then the three-level endpoint could be considered. However, if the intermediate category is clearly closer to one of the other outcomes, it may be desirable to lump these two outcomes together. Similarly, suppose the investigators judge that the new drug need to be equivalent to the old drug with respect to the “success” outcome, and the distinction between a half-way response and complete failure is largely irrelevant to the assessment of the new drug. Then, clearly, this judgment should drive the decision, and the success endpoint should be the primary endpoint.

We believe that the $\rho=.5$ would be a reasonable choice for most applications, especially, since as noted above, if the intermediate outcome is clearly closer to success or to failure, lumping would probably be prudent. However, when using $\rho=.5$, one might consider sensitivity analyses with other values for ρ such as .3 and .7. Certainly, if treatment difference results seem

concordant across a range of these values, then this would provide more assurance about robustness, where any particular choice of p is questionable. However, we do not believe that there is any true value of p per se.

In the setting of the MRSA infection study that motivated this paper, the final assessment will be based on a clinical assessment, so there is inevitably some element of subjectivity. Thus, another consideration in the decision to use a three-level endpoint would be the reliability of all three outcome designations. For example, what if any two assessment evaluators would very likely agree on the outcome assessment if they only needed to choose between “success” and “failure”, but if they had an option of an additional third intermediate category, then their agreement is substantially reduced? In such a scenario, a binary endpoint would probably be preferable and easier to interpret. On the other hand, when the reverse is true, where agreement is better with the three categories, because outcomes naturally fall into three clusters, the three-level endpoint would probably be preferable.

Specification of a non-inferiority margin should be done with the specific trial setting in mind, and the choice of this margin when using the three-level endpoint may be even less straightforward than the selection for a binary endpoint. However, with careful consultation between statisticians and clinical investigators, it can be done. Non-inferiority region plots such as those presented in Figure 1 might be a helpful consulting tool.

An alternative approach would be to choose one of the lumped binary endpoints as the primary endpoint, and if this is satisfied then consider the other lumped binary endpoint as a secondary endpoint. Similarly, these analyses could be co-primary. These approaches would require taking into account the correlations between these analyses. And, if we want to demonstrate non-inferiority on both of these endpoints, we could actually relax the statistical significance levels. That is, this is the opposite of the usual multiple comparisons situation. If we require a statistical result on *both* analyses, this is the reverse of considering a drug a winner when it has a significant result on one of two.

This paper focuses on trials that define differences in proportions as the treatment effect. This is the metric typically used in trials of anti-infective drugs. We note, however, that if the metric of interest was the odds ratio, or relative risk, then the relative sample sizes of the three approaches could be quite different.

In conclusion, the method described in our paper is a very simple procedure, which can easily be employed. While it will not always be a prudent choice, investigators might consider implementing this approach when: a) they wish assurance that the two regimens are sufficiently similar at these three levels of outcome, when the trichotomous categorization more meaningfully represents the continuum of clinical response than a binary outcome, b) outcomes naturally cluster into three categories so that evaluators can provide cleaner and more consistent assessments, or c) sample size calculations suggest it would be an efficient, and low-risk, strategy. When implicitly continuous responses can be more easily classified into three categories than two, and where the middle category can reasonably be assigned a value p , this new method essentially presents a compromise position between either of the binary approaches: using either success alone, or success and intermediate together.

Acknowledgments

The authors wish to thank our colleagues, Dr. Dean Follmann and Dr. Christine Choiu, for useful discussions regarding this manuscript. We also thank the anonymous reviewers for helpful comments that improved the paper.

References

- Agresti, A. Categorical Data Analysis. Wiley; Hoboken: 2002.
- Blackwelder WC. Proving the null hypothesis. *Controlled Clinical Trials* 1982;3:345–353. [PubMed: 7160191]
- Blackwelder, WC. Encyclopedia of Biostatistics. Second Edition. Wiley; West Sussex England: 2005. Equivalence Trials; p. 1735-1740.
- D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* 2003;22:169–186. [PubMed: 12520555]
- Daly JS, Worthington MG, Andrews RJ, Brown RB, Schwartz R, Sexton DJ. Randomized, double-blind trial of cefonicid and nafcillin in the treatment of skin and skin structure infections. *Antimicrobial Agents and Chemotherapy* 1990;34:654–656. [PubMed: 2344171]
- Farrington CP, 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. [PubMed: 2281232]
- Hung HMJ, Wang S-J, Tsong Y, Lawrence J, O'Neill RT. Some fundamental issues with non-inferiority testing in active control trials. *Statistics in Medicine* 2003;22:213–225. [PubMed: 12520558]
- International Conference on Harmonization E-10. Guidance on choice of control group and related design and conduct issues in clinical trials. Food and Drug Administration, DHHS. July;2000 2002.
- Lode H, Magyar P, Muir JF, Loos U, Kleutgens K. Once-daily oral gatifloxacin vs. three-times-daily co-amoxiclav in the treatment of patients with community-acquired pneumonia. *Clinical Microbiology and Infection* 2004;10:512–520. [PubMed: 15191378]
- Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports* 1978;62:1037–1040. [PubMed: 688245]
- Roebuck P, Kuhn 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. [PubMed: 7481195]
- Weins B. Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clinical Trials* 2002;23:2–14. [PubMed: 11852160]

1A

1B

Figure 1.

Regions of non-inferiority for the true experimental group proportions when P_{SC} and P_{IC} are set to specific values and the margin is set to .10, are shown for each endpoint; note that $\rho=.5$. η_R is the region of (P_{SE}, P_{IE}) that meet the non-inferiority standard; η_3 and η_S are likewise. (In Figure 1A, $P_{SC}=.8$ and $P_{IC}=.1$; in Figure 1B, $P_{SC}=.6$ and $P_{IC}=.3$)

Figure 2.

Endpoint with the smallest sample requirement where the margin preserves the same proportion of the benefit of the active control, for $p=.5$. Ω_R is the region of (P_S, P_I) where the response endpoint requires the smallest sample size of the three endpoints; Ω_3 and Ω_S are likewise. Here $\gamma = (P_R - P_R^*) / (P_S - P_S^*)$, where P_R^* is the proportion with response with placebo, and P_S^* is likewise. Alternatively, when the margins are specified such that $\Delta_S = \Delta_R = \Delta_3$, the regions in the $\gamma=1$ plot apply.

Table 1

Relative Sample Sizes of Binary Endpoint to Three-Level Endpoint as a Function of P_S and P_I when Non-inferiority Margin is Same for each Endpoint; n_R is the required sample size for the response endpoint; n_3 and n_S are likewise. (Note that $\rho=.5$.)

P_I	P_S	Relative Sample Size	
		n_S / n_3	n_R / n_3
.1	.2	.98	1.29
	.5	1.12	1.08
	.8	1.56	.88
.4	.1	.82	2.27
	.3	1.40	1.40
	.5	2.27	.82

Table 2

Simulation Study: $\Delta_3=.1$ and $\rho=.5$; sample size computed for 90% power

Case		True Rates				N Per Group	Simulated P(Non-inferiority concluded)
		P _{SC}	P _{TC}	P _{SE}	P _{TE}		
1	$\theta_3=0$ 10,000 replications	.45	.45	.45	.45	225	.900
2		.70	.20	.70	.20	232	.903
3		.20	.70	.20	.70	153	.905
4		.35	.35	.35	.35	341	.898
5		.50	.20	.50	.20	400	.907
6		.20	.50	.20	.50	258	.910
7	$\theta_3=\Delta_3$ 100,000 replications	.45	.45	.35	.45	225	.0249
8		.70	.20	.60	.20	232	.0255
9		.20	.70	.10	.70	153	.0271
10		.35	.35	.25	.35	341	.0255
11		.50	.20	.40	.20	400	.0247
12		.20	.50	.10	.50	258	.0256

Table 3

Three endpoints in results of trial in community-acquired pneumonia (Lode et al. 2004)

	End of Treatment: Observed data		End of Study: Observed data	
	Experimental	Control	Experimental	Control
Cure	123	108	141	133
Improvement	75	78	43	33
Non-Response	30	42	27	45
Total	228	228	211	211
	End of Treatment: Treatment difference result		End of Study: Treatment difference result	
Endpoint	Point estimate	95% CI	Point estimate	95% CI
Success	.066	(-.026, .157)	.038	(-.053, .129)
Response	.053	(-.014, .119)	.085	(.014, .157)
Three-level	.059	(-.008, .127)	.062	(-.011, .135)