Published in final edited form as:

Stat Med. 2021 January 30; 40(2): 327–348. doi:10.1002/sim.8777.

A framework for considering the risk-benefit trade-off in designing noninferiority trials using composite outcome approaches

Grace Montepiedra¹, Ritesh Ramchandani¹, Sachiko Miyahara¹, Soyeon Kim²

¹Harvard T.H. Chan School of Public Health, Boston, MA, USA

²Frontier Science Foundation, Boston, MA, USA

Abstract

When a new treatment regimen is expected to have comparable or slightly worse efficacy to that of the control regimen but has benefits in other domains such as safety and tolerability, a non-inferiority (NI) trial may be appropriate but is fraught with difficulty in justifying an acceptable NI margin that is based on both clinical and statistical input. To overcome this, we propose to utilize composite risk-benefit outcomes that combine elements from domains of importance (e.g. efficacy, safety, and tolerability). The composite outcome itself may be analyzed using a superiority framework, or it can be used as a tool at the design stage of a NI trial for selecting an NI margin for efficacy that balances changes in risks and benefits. In the latter case, the choice of NI margin may be based on a novel quantity called the maximum allowable decrease in efficacy (MADE), defined as the marginal difference in efficacy between arms that would yield a null treatment effect for the composite outcome given an assumed distribution for the composite outcome. We observe that MADE: (1) is larger when the safety improvement for the experimental arm is larger, (2) depends on the association between the efficacy and safety outcomes, and (3) depends on the control arm efficacy rate. We use a numerical example for power comparisons between a superiority test for the composite outcome versus a non-inferiority test for efficacy using the MADE as the NI margin, and apply the methods to a TB treatment trial.

Keywords

risk-benefit; non-inferiority margin; clinical trial; composite outcome; trial design; tuberculosis

1. Introduction

For diseases for which a known efficacious treatment is available, studies evaluating a new treatment will typically compare the new treatment to an active control. Researchers will use a superiority comparison if the experimental treatment is expected to have better efficacy than the control but will use a non-inferiority comparison for efficacy when the control treatment is known to be efficacious but there is an anticipated benefit, in a domain other

than efficacy of the experimental treatment over the control treatment. The benefit, e.g., a clinically important improvement in safety, may offset a potential modest decrease in efficacy.

One approach is to compare efficacy and safety separately, by performing two simultaneous tests in the primary analysis: a non-inferiority test on efficacy and a superiority test on safety. The safety benefit of the experimental treatment over the control treatment allows the experimental treatment arm to have an efficacy rate that is lower than that of the control arm, which is reflected by the selected non-inferiority margin. In many non-inferiority trials, the efficacy comparison is the sole primary objective, and the safety comparison is relegated to be a secondary objective.

Non-inferiority trials are relevant to evaluation of treatment regimens for tuberculosis (TB), which remains the most serious infectious disease globally, contributing substantially to worldwide mortality and morbidity. For the past decade, there has been a resurgence of confirmatory trials evaluating regimens for the treatment and prevention of tuberculosis, in both drug-susceptible (DS)-TB and multidrug-resistant (MDR)-TB. Results of randomized TB trials have been, and continue to be, of highest public health importance, confirmed by their publication in high impact medical journals. The New England Journal of Medicine has published results of almost all of these randomized studies^{1–12}, and the vast majority of these have been non-inferiority trials^{3,5–7,10–12}. Most randomized TB trials that are ongoing or in development are non-inferiority trials. Non-inferiority trials have been of great interest to the TB research community because of the well-known toxicities and tolerability issues associated with some drugs used to treat TB, including irreversible hearing loss and fulminant hepatotoxicity, which suggest a need for improved safety of regimens to treat and prevent TB. A regimen offering such a benefit might be preferred over the currently recommended regimen even with some loss of efficacy. The key issue in the design of non-inferiority trials for treatment/prevention of TB is the determination of the non-inferiority margin (NIM). This has been, and continues to be, a controversial issue, and conclusions from completed TB trials continue to be debated precisely because of the choice of this NIM. In their recent review paper, Phillips et al¹³ noted that “[t]he choice of the margin in trials of TB treatment regimens continues to be a major discussion issue.” Olliaro et al¹⁴ recently provided a review of non-inferiority TB treatment trials, with an extensive discussion of the issues encountered with the choice of the NIM in these studies.

There is little practical guidance in the determination of the NIM, either from a statistical or clinical perspective. The FDA guidance¹⁵ is limited to suggesting that the margin should be no larger than the treatment effect (M1) of the active control relative to placebo, and that the margin should reflect the loss of effect that is considered clinically acceptable. Among published TB trials, Sterling et al³ provided the most extensive discussion and clearest explanation of how the choice of the NIM preserves a large percentage (79%) of M1, while others use this as rationale but lacked a clear explanation of how the NIM was derived.

The FDA guidance¹⁵ also states that “a larger value for [the margin] may be justified clinically if the test drug were shown to have an important advantage, (e.g. on safety or on a secondary endpoint).” This language clearly indicates that the non-inferiority margin is tied

to a risk-benefit trade-off between the primary outcome, usually efficacy, and other secondary outcomes such as safety. Phillips et al¹³ asserted that “consideration of the expected benefits of the intervention shapes the final choice of margin.....Anticipated reduction in toxicity is [one] factor that may influence the choice of margin; whether this is a consequence of the new treatment or not cannot be known until the trial has been completed. Collection of good safety data is critical to properly weigh the risks and benefits of an intervention.” In some of these TB trials, the “expected benefit” of the experimental regimen was explicitly used as a rationale for the non-inferiority trial design. However, given the absence of a framework (both clinical and statistical) to guide quantification of the trade-off and thus this “expected benefit”, there has been no discussion of how this “expected benefit” informed the NIM. Implicitly, a study can be designed as a non-inferiority trial only if the experimental regimen offsets some loss of efficacy. Ollario et al¹⁴ argued for the need to establish a non-inferiority margin that is statistically and clinically “meaningful”, by capturing the trade-off between gains and losses in a quantifiable manner.

Many methods have been proposed in the literature to combine multiple outcomes and analyze the joint outcome to evaluate the treatment that is more favorable overall. The idea of combining efficacy and safety into a ranked outcome was introduced by Chuang-Stein and Mohberg¹⁶. Generalized prioritized outcome approaches were discussed much later on by Buyse¹⁷, who formalized a treatment comparison parameter called the proportion in favor of treatment, and Pocock et al¹⁸, who introduced the win ratio parameter. Evans et al¹⁹ defined the Desirability of Outcome Ranking (DOOR) outcome in the context of antibiotic stewardship trials. Another approach assigns weights or partial scores to components (or combination of these), and incorporate them in a composite outcome^{20–23}. These approaches can obviate the need to perform a non-inferiority comparison since the risk-benefit tradeoff is embedded into the single outcome. There has been an increasing interest within the TB clinical trials research community to consider using the DOOR-type outcomes as an alternative to, or in tandem with, performing non-inferiority comparison. This includes comparisons of these two approaches during study designs, with the intent to address the trade-off between risks versus benefits of regimens with these approaches. Miyahara et al²⁴ were the first to apply such risk-benefit composite outcomes to TB data from a recently published non-inferiority study that showed a 1-month regimen was non-inferior to a 9-month regimen for the prevention of TB.

To this end, we provide a unified framework that links a superiority comparison using a DOOR-type outcome with an equivalently formulated non-inferiority comparison by appropriate choice of the NIM. We propose to capitalize on the trade-off between efficacy and safety that is implicit in a comparison of regimens with respect to a DOOR-type outcome, by quantifying this trade-off, which can inform the determination of the NIM in a non-inferiority trial. In Section 2 of this paper, we formalize a risk-benefit composite ranked outcome (DOOR) combining two binary outcomes in a clinical trial setting. We then set up the framework for a test of superiority of the experimental treatment over the control treatment based on the effect size, a parameter that compares the two arms with respect to this risk-benefit outcome. In Section 3, we show that the effect size is a monotonic increasing function of the experimental arm efficacy rate, for fixed safety rates in both arms and fixed efficacy rate in the control arm. Based on this property, we introduce a novel

concept called the maximum allowable decrease in efficacy (MADE) in the experimental treatment, a value that quantifies the trade-off in efficacy for a test treatment with a better safety profile than the control treatment. In the setting of a non-inferiority trial, we propose to use MADE as a formal statistical tool to inform the selection of the NIM for the primary outcome that balances risk and benefit. Section 4 provides the corresponding framework when the risk-benefit composite outcome is a weighted combination of the two binary outcomes. For all risk-benefit outcomes defined in this paper, we illustrate how MADE quantifies the inherent trade-off between risk and benefit. Finally, in a variety of simulated scenarios motivated by a pediatric MDR-TB treatment trial presented in Section 5, we observe a modest to substantial increase in power for a non-inferiority approach to establishing efficacy compared to using a superiority approach for the risk-benefit outcome, while both approaches show type I error rates similar to the nominal level of significance used. We follow with discussion of when the validity of choice of the NIM for a non-inferiority trial may be in question in cases of large deviations from assumptions that informed the value of MADE.

2. The DOOR Approach

The justification for conducting a non-inferiority trial is often to declare superiority of the regimen when considering the totality of its effect with respect to important factors, e.g., a combination of efficacy, safety, and tolerability. An alternative approach uses a composite outcome that ranks patients' experience with respect to clinical benefit (a favorable outcome or not), presence or absence of any treatment-limiting adverse events, and other factors, more generally called a DOOR outcome^{19, 24, 25}.

In this paper, we consider two binary domains of interest, efficacy and safety, which are the most commonly studied primary outcomes in randomized confirmatory TB trials. For example, in a non-inferiority trial comparing a four-month shortened regimen to the standard six-month regimen for the treatment of DS-TB, Gillespie et al⁷ had two primary binary outcomes. The efficacy outcome was classified as treatment success or treatment failure, where treatment success is defined as the participant having bacteriological cure and no relapse within one year of completion of therapy. The safety outcome was an indicator of whether or not the participant experienced a Grade 3 or 4 adverse event (using a Modified Division of Acquired Immunodeficiency Syndrome/National Institute of Allergy and Infectious Diseases [DAIDS/NIAID] Scale of Adverse Event Reporting). If efficacy is given higher priority over safety, a reasonable ranking of the joint efficacy/safety outcomes is shown in Table 1. If safety has priority over efficacy, the middle rankings would be reversed.

Let C , T be the DOOR outcome of a randomly selected patient in the control arm C and experimental arm T , respectively. Let $\pi_C(j)$ and $\pi_T(j)$ be the probability of a DOOR outcome j for a patient in arm C and arm T , respectively. C and T follow a multinomial distribution corresponding to the joint probabilities of each of the four combinations of efficacy and safety events in Table 1. The primary comparison of arms will be the probability that a patient will have a higher ranking (that is, better outcome) on treatment T versus the treatment C , which occurs when

$$P(T > C) - P(C > T) > 0 \quad (1)$$

where

$$P(T > C) = \pi_T(4) \times [\pi_C(3) + \pi_C(2) + \pi_C(1)] + \pi_T(3) \times [\pi_C(2) + \pi_C(1)] + \pi_T(2) \times \pi_C(1) \quad (2)$$

and

$$P(C > T) = \pi_C(4) \times [\pi_T(3) + \pi_T(2) + \pi_T(1)] + \pi_C(3) \times [\pi_T(2) + \pi_T(1)] + \pi_C(2) \times \pi_T(1) \quad (3)$$

If (1) does not hold, then the experimental arm is the same as, or worse than, the control arm with respect to the composite outcome. Because the outcome incorporates both efficacy and safety, using the DOOR outcome for comparison of treatment arms enables a single hypothesis test of superiority of the experimental arm in which we test the upper-tailed alternative hypothesis described by the inequality (1). The hypothesis in expression (1) is an example of what has been called generalized or nonparametric Behrens-Fisher problem^{26, 27}.

To illustrate, assume that the probability distributions of the DOOR outcome for treatments C and T are as described in Table 2a. From expressions (2) and (3) it can be shown that $P(T > C) = 27.5\%$ and $P(C > T) = 14.1\%$, and from (1) we have $P(T > C) - P(C > T) = 27.5\% - 14.1\% = 13.4\%$. We can say that treatment T is superior to treatment C with respect to the DOOR outcome and that treatment T “won” over treatment C by an *effect size* of 13.4%. This effect size parameter is similar to general measures of the difference between treatments such as the “proportion in favor of treatment”² or the “global treatment effect”^{28, 29}.

The joint probabilities in Table 2a can equivalently be represented by the marginal probabilities for efficacy and safety in each arm, and a parameter describing the association between efficacy and safety in each arm³⁰. Let us denote the marginal probability of a favorable efficacy outcome under treatment C and treatment T by the quantities P_{CE} and P_{TE} , respectively, so that

$$P_{KE} = \pi_K(3) + \pi_K(4), \quad K = T, C. \quad (4)$$

Similarly, we denote the marginal probability of a participant experiencing no treatment-limiting adverse events under treatment C and treatment T by P_{CS} and P_{TS} , respectively, so that

$$P_{KS} = \pi_K(2) + \pi_K(4), \quad K = T, C. \quad (5)$$

The association between efficacy and safety in each arm can be parameterized in at least two ways; one can use: (a) the odds ratios under treatment C and treatment T denoted by R_C and R_T , respectively, where:

$$R_K = [\pi_K(1) \times \pi_K(4)] / [\pi_K(2) \times \pi_K(3)], \quad K = T, C \quad (6)$$

or (b) the correlation coefficients under treatment C and treatment T denoted by ρ_C and ρ_T , where

$$\rho_K = [\pi_K(4) - P_{KE} \times P_{KS}] / [P_{KE}(1 - P_{KE})P_{KS}(1 - P_{KS})]^{1/2}, \quad K = T, C.$$

Since odds ratios are commonly used to describe associations between binary variables, the odds ratios R_C and R_T will be used throughout the paper to represent the association between efficacy and safety for the two treatment arms. When efficacy and safety are independent in each arm (e.g., $R_C = R_T = R = 1$), the multinomial probabilities associated with random variables C and T can be obtained from the marginal probabilities for efficacy and safety as follows:

$$\pi_K(1) = [1 - P_{KE}] \times [1 - P_{KS}], \quad (7)$$

$$\pi_K(2) = [1 - P_{KE}] \times P_{KS},$$

$$\pi_K(3) = P_{KE} \times [1 - P_{KS}],$$

$$\pi_K(4) = P_{KE} \times P_{KS}, \quad K = T, C.$$

The joint distribution for efficacy and safety for each of the arms shown in Table 2a can be represented by the marginal probabilities and odds ratio in Table 2b. For this scenario, efficacy and safety outcomes are independent ($R_C = R_T = 1$) in each arm.

However, it may be more plausible to assume that efficacy and safety are associated, with patients who do not experience a treatment-limiting adverse event being more likely to have a favorable treatment outcome ($R_C > 1$ and $R_T > 1$). For the example where Table 2b represents the true marginal probabilities for efficacy and safety in treatment C and T, Figure 1 portrays the trajectory of effect size, $P(T > C) - P(C > T)$, as the odds ratio increases from 0 to 10. In this scenario, the true effect size is always positive regardless of the direction of dependence between efficacy and safety, implying that treatment T is superior; however, the magnitude of the effect size decreases as R increases.

3. The Risk-Benefit Ranked Outcome (DOOR) and the Efficacy/Safety Tradeoff

The appeal of the DOOR outcome is that it inherently accounts for the trade-off between efficacy and safety, through the rankings. For simplicity of discussion, let us assume that efficacy and safety outcomes are independent ($R_C = R_T = 1$) in each treatment arm. Under

this assumption, and using the DOOR outcome defined in Table 1, it can be shown that treatment arm T is better than treatment arm C in the sense of inequality (1) if and only if the following inequality is satisfied by the marginal probabilities of the arms:

$$\delta_E < -\delta_S \times w_E, \quad (8)$$

where $\delta_E = P_{CE} - P_{TE}$, $\delta_S = P_{CS} - P_{TS}$, and $w_E = P_{CE} \times P_{TE} + [(1 - P_{CE})(1 - P_{TE})]$.

Refer to the Appendix for a proof of this equivalence. Since δ_E and δ_S are readily interpretable as the decrease in favorable efficacy and safety probabilities of treatment T vs treatment C, respectively, expression (8) shows that superiority of experimental arm T over control arm C allows the scenario where the control arm can have better efficacy than the experimental arm, as long as it is less than the difference in safety probabilities between the two treatment arms multiplied by the weighting factor w_E , which is a function of the efficacy rates of the treatment arms. There is, thus, an efficacy/safety trade-off implicit in a superiority comparison using the DOOR outcome. Rauch³¹ provides a generalized expression for (8) when there are multiple independent binary components in the composite outcome.

This trade-off gives us the insight that we can select a non-inferiority margin for the efficacy outcome based on the DOOR outcome. While the treatment effect of T vs. C is increasing as a function of the marginal probability P_{TE} with fixed P_{TS} , P_{CE} , P_{CS} , and fixed dependence of efficacy and safety for T and C, the non-inferiority margin can be chosen as the value of δ_E that makes the overall treatment effect equal to 0.

Theorem 1:

Let P_{CE} , P_{CS} , P_{TS} , and the dependence parameters R_C and R_T , be fixed values. Then only one of the following is true:

- i. There exists a unique $P_{TE} = P_{TE}^*$ that satisfies $P(T>C) - P(C>T) = 0$,
 $P(T>C) - P(C>T) > 0$ when $P_{TE} > P_{TE}^*$, and
 $P(T>C) - P(C>T) < 0$ when $P_{TE} < P_{TE}^*$.
- ii. $P(T>C) - P(C>T) > 0$ for every P_{TE} in $(0,1)$
- iii. $P(T>C) - P(C>T) < 0$ for every P_{TE} in $(0,1)$

Proof: From Lemma 3 in the Appendix, $P(T>C) - P(C>T)$ is monotone increasing in P_{TE} . The result immediately follows.

Lemma 3 asserts that as the efficacy of treatment T increases, the superiority of the experimental treatment over the control treatment increases with respect to the DOOR outcome. Keeping all other parameters fixed, if statement (ii) of Theorem 1 is true, then treatment T is always superior to treatment C regardless of the value of P_{TE} . If statement (iii) is true, then treatment T is always inferior to treatment C. If (i) is true, which is the case when $P(T>C) - P(C>T)$ crosses 0, then there is a unique minimum value for P_{TE} , denoted by

P_{TE}^* , where superiority of the experimental treatment arm holds. If (i) is true, as long as treatment T has efficacy that is above P_{TE}^* then it is superior to treatment C.

In general, let $\delta_E^* = P_{CE} - P_{TE}^*$. We can call the quantity δ_E^* the **maximum allowable decrease in efficacy (MADE)** for treatment T to be superior to treatment C under fixed efficacy rate for treatment C, fixed safety rates for treatments T and C, and fixed odds ratios R_T and R_C for treatments T and C.

The scenario where the function $P(T>C) - P(C>T)$ is always negative with respect to P_{TE} occurs when, for example, the active control has a high efficacy rate, say 95%, and its safety profile is much better than treatment T (for example, $P_{TS} = 20\%$ and $P_{CS} = 80\%$). In this case, treatment T would not be a logical candidate to compete against treatment C given its worse safety profile. The scenario where the function is always negative with respect to P_{TE} occurs when, for example, the active control has a low efficacy rate, say 30%, and treatment T is anticipated to have a much better safety profile (for example, $P_{TS} = 90\%$ and $P_{CS} = 10\%$). This case implies that the trade-off in efficacy due to an anticipated high safety benefit allows treatment T to have no efficacy at all - that is, P_{TE} can be zero and yet be superior to the active control in the sense of the DOOR outcome. In this case, if a non-inferiority trial were to be designed, the NIM could be fully based on preservation of a substantial fraction of M1, the improvement of the active control over the placebo. The more interesting case is when the function $P(T>C) - P(C>T)$ crosses zero, which is when MADE can be useful for informing the choice of NIM; it occurs when evaluation of treatment T largely relies on its anticipated safety benefit to justify any loss in efficacy and is the main rationale for using a non-inferiority study design.

Theorem 2:

Suppose that the binary efficacy and safety outcomes are independent for both treatment arms and statement (i) in Theorem 1 is true. Given P_{CE} and $\delta_S = P_{CS} - P_{TS}$, the **maximum allowable decrease in efficacy (MADE)** for treatment T to be superior to treatment C with respect to the DOOR outcome is $\delta_E^* = P_{CE} - P_{TE}^*$, where

$$P_{TE}^* = \{P_{CE} - \delta_S \times [P_{CE} - 1]\} / \{1 - \delta_S \times [2 \times P_{CE} - 1]\}, \quad (9)$$

whenever the denominator in (9) is not equal to zero, that is, whenever $\delta_S \times [2 \times P_{CE} - 1] \neq 1$.

Proof: Since (1) is equivalent to (8) when $R_C = R_T = 1$, and with P_{TE}^* defined in Theorem 1 (i), solving for P_{TE} in the equality $\delta_E = -\delta_S \times w_E$ will give P_{TE}^* . After some algebraic manipulation, it can be shown that the solution is given by (9).

To illustrate, consider the case when the control arm has efficacy (P_{CE}) of 85% and efficacy and safety are independent. Table 3 provides values of MADE computed using Theorem 2 under varying true safety rate differences between treatment arms (first column). When $-\delta_S = P_{TS} - P_{CS}$ is 5%, the DOOR effect size decreases from 5.80% to -1.45% as P_{TE} decreases from 87% to 80%. The experimental arm T retains its superiority over the control arm C with respect to the DOOR outcome in Table 1 (i.e., effect size > 0) when $P_{TE} > 81.4\%$. The effect size reaches 0 when P_{TE} decreases to 81.4%. Expression (8) shows that the superiority

depends on the safety rates only through their difference; note that this does not necessarily hold if efficacy and safety are dependent in either treatment arm (i.e., $R_C \neq 1$ or $R_T \neq 1$). This highlights an important benefit of the risk-benefit approach in the case that the safety rate of a new regimen is not accurately known at the design stage of a trial because, as Miyahara et al²⁴ point out, it allows gains in one dimension to offset losses in another. In this example, if the true difference in safety rates were 5%, then experimental treatment T can have an efficacy rate up to 3.6% lower and maintain its overall superiority over control arm C. The maximum allowable decrease in efficacy (MADE) in this example is 3.6%. Assuming the rankings defined by the DOOR outcome in Table 1 capture the risk-benefit trade-off, if the true difference in safety rates were 15%, then $MADE = 10.12\%$. Clearly, this allowable loss in efficacy depends on the magnitude of difference in safety rates. If a trial is designed to compare superiority of the experimental regimen T over the control regimen C with respect to the DOOR outcome, there is no need to identify this allowable “margin” since the trade-off is inherent to the approach, unlike non-inferiority trials where this can be a controversial design issue.

Table 4 provides values of MADE assuming that the control regimen has efficacy rate (P_{CE}) of 85% and safety rate (P_{CS}) of 80%, under varying true safety rate differences between treatment arms (first column), and this time allowing for dependence of safety and efficacy, under various values of $R_C = R_T = R$ (second column) which allow for positive, negative, or no correlation. Since there is no closed form solution when $R \neq 1$, the MADE was determined numerically using a root-finding algorithm. Note that if $R = 1$ then this trade-off depends not only on the true difference in safety rates but also on the true safety rate in the control arm, P_{CS} . Assuming that the true safety rate difference between arms is 15%, MADE varies from a minimum of 6.4% (as $R \rightarrow \infty$) to a maximum of 10.6% (as $R \rightarrow 0$). For example, if safety and efficacy are strongly positively correlated with $R = 10$, then $MADE = 8.8\%$, which means that experimental arm T maintains its superiority over control arm C while it has an efficacy rate that is $> 76.2\%$. If safety and efficacy are independent ($R = 1$), then $MADE = 10.1\%$ and experimental arm T can have an efficacy rate as low as 74.9%. A similar trend of decreasing MADE as R increases is also observed under the scenario when the true safety rate difference between treatment and control arms is 10%. Interestingly, the trend reverses if the true safety rate difference between treatment arms is 5%, with lowest MADE of 3.5% occurring under condition of positive association between safety and efficacy and highest MADE of 4.7% under condition of negative association of these outcomes. From Figure 2, which provides a visual summary of Table 4, it is clear that MADE can be significantly affected by $-\delta_S$, the magnitude of gain in safety with treatment T, but is fairly robust to varying values of R .

The concept of MADE sheds light on the nature and magnitude of the allowable trade-off between efficacy and safety when using a risk-benefit outcome such as the DOOR outcome defined in Table 1, which incorporates preferences of efficacy vs safety through the rankings. Moreover, quantifying this trade-off can be beneficial when designing a non-inferiority trial. Under the same assumptions and independence described in Table 4, if the main rationale for prescribing a new regimen with a lower favorable outcome rate is the expectation of a higher safety rate compared to the standard treatment by a magnitude of 15%, one can propose a non-inferiority trial that uses the MADE value of 10.1% as the non-

inferiority margin. If there are other anticipated risk reductions or benefits that treatment regimen T has over the control regimen C that have not been incorporated in the ranking, then the NI margin can be larger than 10.1%. If it is possible to incorporate all relevant factors associated with treatments T and C into the DOOR outcome and derive the corresponding MADE parameter, then the researcher will have a comprehensive framework for quantifying the acceptable trade-off and therefore an enhanced tool to inform the NI margin.

4. The Risk-Benefit Weighted Outcomes and the Efficacy/Safety Tradeoff

A limitation of using a risk-benefit ranked outcome is the inability to incorporate the magnitude of importance of one outcome over the other. For instance, a patient, caregiver or health care provider may believe that experiencing a treatment-limiting adverse event and having an unfavorable outcome is substantially worse than having an unfavorable outcome without experiencing a treatment-limiting adverse event. In this case, it might be important to incorporate the extent of how much worse the outcome is in the assessment of a new treatment or regimen. A composite outcome based on weights²³ or partial scores²² placed on each of the four possible outcomes might be used to overcome this inadequacy.

4.1 Weighted Composite Outcome

Let C_W and T_W have values w_i , $\{i = 1, 2, 3, 4\}$, where $0 < w_1 = w_2 = w_3 = w_4 < 1$ and $\sum_{i=1}^4 w_i = 1$, that correspond to the weight scores for each of the DOOR outcomes in Table 1. We define C_W and T_W as the random variables corresponding to the weighted composite outcome for randomly selected participants from arm C and arm T, respectively. Then C_W and T_W follow the same multinomial probability distribution as random variables C and T, which were both defined in Section 2. That is,

$$\begin{aligned} P(C_W = w_i) &= \pi_C(i) \text{ and} \\ P(T_W = w_i) &= \pi_T(i) \text{ for } i = 1, 2, 3, 4. \end{aligned} \quad (10)$$

Let

$$\begin{aligned} \mu_C &= E[C_W] = \sum_{i=1}^4 w_i \times \pi_C(i) \text{ and} \\ \mu_T &= E[T_W] = \sum_{i=1}^4 w_i \times \pi_T(i) \end{aligned} \quad (11)$$

be the average weight score for the two treatment arms. A comparison of the totality of the treatment experiences could be based on the difference between these average weight scores, $\mu_T - \mu_C$. We can therefore say that treatment arm T is superior to control arm C if

$$\mu_T - \mu_C > 0. \quad (12)$$

Theorem 3:

Given fixed values for P_{CE} , P_{CS} , P_{TS} , and fixed values for the dependent parameters, then only one of the following is true:

- i. There exists a unique $P_{TE} = P_{TE}^*$ that satisfies $\mu_T - \mu_C = 0$,
 $\mu_T - \mu_C > 0$ when $P_{TE} > P_{TE}^*$, and
 $\mu_T - \mu_C < 0$ when $P_{TE} < P_{TE}^*$
- ii. $\mu_T - \mu_C > 0$ for every P_{TE} in $(0,1)$
- iii. $\mu_T - \mu_C < 0$ for every P_{TE} in $(0,1)$

Proof: From Lemma 4 in the Appendix, $\mu_T - \mu_C$ is monotone increasing in P_{TE} . The result immediately follows.

Analogous to Lemma 3, Lemma 4 asserts that as its efficacy rate increases, then the superiority of the experimental treatment over the control treatment increases with respect to the weighted composite outcome. Keeping all other parameters fixed, Theorem 3 directly follows from this assertion, and states that when (i) is true then there is a unique minimum value for P_{TE} , denoted by P_{TE}^* , above which superiority of the experimental treatment arm holds. If (i) is true, as long as treatment T has efficacy that is above P_{TE}^* then it is superior to treatment C.

Theorem 4:

Suppose that binary efficacy and safety outcomes are independent for both treatment arms ($R_C=R_T=R=1$) and (i) in Theorem 3 is true. Given $w = \{w_1, w_2, w_3, w_4\}$, P_{CE} , P_{CS} and P_{TS} , then the **maximum allowable decrease in efficacy (MADE)** for treatment T to be superior to treatment C with respect to the weighted composite outcome is $\delta_E^* = P_{CE} - P_{TE}^*$, where

$$P_{TE}^* = \{[k_1(w) + k_2(w) \times P_{CS}] \times P_{CE} + k_3(w) \times \delta_S\} / \{k_1(w) + k_2(w) \times P_{TS}\} \quad (13)$$

with $k_1(\underline{w}) = w_3 - w_1$, $k_2(\underline{w}) = w_1 - w_2 - w_3 + w_4$, $k_3(\underline{w}) = w_2 - w_1$ and $\delta_S = P_{CS} - P_{TS}$. This quantity exists while the denominator of (13) is non-zero.

Proof: Since $R = 1$, substitute (7) into (11) to get

$$\mu_T - \mu_C = k_1(\underline{w}) \times [P_{TE} - P_{CE}] + k_3(\underline{w}) \times [P_{TS} - P_{CS}] + k_2(\underline{w}) \times [P_{TE} \times P_{TS} - P_{CE} \times P_{CS}]. \quad (14)$$

With P_{TE}^* defined in Theorem 3 (i), equating the right hand side of (17) to 0 and solving for P_{TE} will give P_{TE}^* . After some algebraic manipulation, it can be shown that the solution is given by (13).

4.2 Weighted Component Outcome

Instead of assigning weights to the four possible outcomes in Table 1, weight scores can be assigned to each of the two binary component outcomes for efficacy and safety marginally. A weight score ω_E can be assigned to the efficacy outcome and a weight score of $1 - \omega_E$ can be assigned to the safety outcome, where $0 < \omega_E < 1$. A higher weight implies a higher preference or level of importance given to the component outcome. Treatment T would then be better than treatment C if and only if

$$\omega_E \times P_{CE} + (1 - \omega_E) \times P_{CS} < \omega_E \times P_{TE} + (1 - \omega_E) \times P_{TS}, \quad (15)$$

which is equivalent to

$$[P_{CE} - P_{TE}] < k(\omega_E) \times [P_{TS} - P_{CS}], \quad (16)$$

where $k(\omega_E) = (1 - \omega_E)/\omega_E$. Based on this weighted component outcome, MADE can readily be seen as the right hand side of inequality (16), $k(\omega_E) \times [P_{CS} - P_{TS}]$.

Table 5 shows the corresponding values of MADE for treatment T to be superior to treatment C in the sense of (15), under varied weights assigned to the efficacy component and under assumptions of 5%, 10% and 15% safety rate differences between arms. For example, suppose that efficacy is given a weight of 0.6 (which means safety is given a weight of 0.4). Assuming that the improvement in safety for treatment T is 15%, then the maximum decrease in efficacy for which treatment T is superior in the sense of (15) is determined by the MADE value of 10.05%. Note that no assumptions on the control arm efficacy and safety rates, or on the magnitude and direction of dependence between efficacy and safety, are required to determine the MADE.

5. Comparison of risk-benefit outcome versus non-inferiority approach in designing a pediatric MDR-TB trial: A numerical study

Given the toxicity and poor tolerability associated with many of the drugs used in standard regimens for treating MDR-TB, an alternative regimen that may have small improvements in, or slightly worse, efficacy may have a more favorable risk-benefit profile if it is less toxic and more tolerable. This is even more relevant in consideration of treatment for children with MDR-TB, where a higher proportion of favorable treatment outcomes have been observed compared to adults with the disease, and children are more severely debilitated by the harmful side effects of several of the drugs used to treat MDR-TB³². For example, aminoglycosides and cyclic peptides are painful because they are administered by intramuscular injections and may cause permanent hearing loss, which can have repercussions on neurocognitive development and quality of life. Thus, Seddon et al³² made the case for the need for pediatric MDR-TB treatment efficacy trials designed for non-inferiority. Kim et al³³ also discussed the value of using a non-inferiority efficacy trial to “identify regimens that are expected to be better or not unacceptably worse than the control and, because of the potential improvements in risk-benefit profile, may be useful in evaluating a shortened, all-oral MDR-TB regimen in children.” However, the choice of the non-inferiority margin continues to be a major design issue, not only in pediatric trials but in adult trials as well³³. Several of the current adult MDR-TB non-inferiority trials have used 10% as the margin (for example, the STREAM trial¹¹), however, there is lack of formal justification on the appropriateness of this quantity for these trials.

In this numerical study, we evaluate the operating characteristics informed by a hypothetical pediatric MDR-TB trial of a novel all-oral shortened treatment regimen, described in Kim et al³³, using the DOOR approach. We compare them to those of a non-inferiority design with

NIM that is based on the value of MADE derived from the same parametric assumptions used in the DOOR approach. We assume that the current treatment will be used in the control arm and is the 9–12 month Bangladesh regimen that received conditional recommendation from the World Health Organization³⁴. This regimen has good efficacy but has an unacceptable safety profile; efficacy is measured by the proportion of patients who have a favorable outcome and safety by the proportion of patients who do not experience serious adverse events, with rates of $P_{CE} = 85\%$ and $P_{CS} = 80\%$, respectively^{35, 36, 37, 38}. Further it is believed that an all-oral 6-month regimen being evaluated is expected to have slightly higher, similar or slightly worse efficacy (P_{TE}) but is also expected to have a much better safety profile, with the proportion of participants without safety events expected to be $P_{TS} = 95\%$ or higher (that is, assume $-\delta_S = P_{TS} - P_{CS} = 15\%$). A marked improvement in safety is anticipated due to the replacement in the Bangladesh regimen of the daily intramuscular injection of an aminoglycoside, known to have a side effect of hearing loss that may be permanent, with a safer drug taken orally, as well the shortening of treatment duration by 3–6 months.

With the expectation that almost all participants will complete the study, a Wilcoxon rank-sum test for superiority of the experimental treatment with respect to the DOOR outcome defined in Table 1, and in the sense of expression (1), is reasonably adequate. Hence, for each simulation, a one-sided Wilcoxon rank-sum test for superiority of the experimental treatment was performed, using a 2.5% level of significance. For each set of a priori design assumptions, 10,000 simulations were run to estimate power and type I error with a selected sample size of $n = 200, 400$ and 1300 per treatment arm under the scenarios when the assumed improvement in safety, $-\delta_S = P_{TS} - P_{CS}$, is 15%, 10% and 5%.

Table 6.1 shows the power to detect superiority of the experimental arm when it is slightly less efficacious ($P_{TE} = 83\%$) compared to the control arm, but has better safety profile. When $-\delta_S = 15\%$ and with 200 participants in each treatment arm, power decreases from 63% to 18% as R ranges from 0 to infinity. It is expected that safety and efficacy are independent or positively correlated; for $1 \leq R \leq 10$ the plausible power of the test ranges from 53% to 35%. When the experimental treatment is slightly more efficacious (Table 6.2, $P_{TE} = 87\%$), power is reasonably good, ranging from 87% to 72% when $1 \leq R \leq 10$; and when efficacy rates of both arms are the same (Table 6.3), power is good to fair, ranging from 72% to 54% under these plausible values of R . When designing a study in practice, it would be important to simulate the power under a variety of scenarios for the parameters in order to choose a sample size that is robust to plausible deviations in the design assumptions.

Alternatively, one can perform a non-inferiority trial for efficacy with MADE used as the non-inferiority margin (NIM). For comparability with the DOOR outcome-based superiority approach detailed above, a two-sided 95% confidence interval for $\delta_E = P_{CE} - P_{TE}$ was computed in each simulation, and a conclusion of non-inferiority of the experimental treatment was made whenever the upper bound of the confidence interval was less than $NIM = MADE$. In the simulations described above, we estimated power to conclude non-inferiority using the MADE as NIM. As seen in Tables 6.1-6.3, using a non-inferiority test for efficacy will have at least as good power as the superiority test with respect to the DOOR outcome, with gains in power generally more pronounced when $R \leq 1$. We also considered

the case when efficacy and safety are co-primary endpoints, with both a test for non-inferiority of the experimental arm with respect to efficacy and a test for superiority of the experimental arm with respect to safety performed so that the overall type I error is preserved at a nominal rate of 2.5%. Compared to a single primary test for non-inferior efficacy, the joint test results in slightly lower power, as anticipated (see last column of Tables 6.1-6.3); however, the power of the joint test is generally still as good as, or better than, testing for superiority using the DOOR approach. Expectedly, as the difference in the safety profiles of the two arms decreases, the MADE value decreases, and power of both the DOOR outcome-based superiority approach and the non-inferiority approach decreases, requiring a larger sample size for adequate power using either approach.

Table 7. shows the corresponding estimated true type I errors, with same simulated data used across all three approaches. Both the Wilcoxon rank sum test on the DOOR outcome and the non-inferiority test for efficacy achieve close to the nominal type I error rate of 2.5% under all scenarios considered, although the non-inferiority test (with or without safety as a co-primary outcome) appears to have a type I error rate that is consistently conservative.

6. Discussion

We have proposed a framework for considering the risk-benefit trade-off when evaluating a new treatment regimen. We describe flexible approaches to developing a risk-benefit outcome, based on rankings, weight scores, or weighted components, the use of which transforms the research question to one of superiority. A notable advantage of the superiority analysis using the composite risk-benefit outcome is that the trade-off is inherent in the definition of the outcome. The composite risk-benefit outcome can be used in a clinical trial to compare arms to assess superiority in the totality of the effect of treatments on a patient, accounting for differing importance of outcome components. This approach matches the overall objectives of a study evaluating a regimen when its benefits, e.g., averting toxicity, improved tolerability, increased simplicity, are expected to justify similar or slightly decreased efficacy with the standard treatment and is therefore superior when viewed in its totality of effect. In a situation where the magnitude of the safety advantage of a regimen may be unknown, a method that takes into account both efficacy and safety may be desired, allowing gains in one dimension to offset losses in another.

Trialists and regulatory agencies agree that determining the NIM requires a more rigorous approach and that both statistical reasoning and clinical judgement are important (D'Agostino et al³⁹, ICH, EMEA⁴⁰, FDA¹⁵). The ICH E10 guidance suggests that the NIM “cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with placebo in the setting of the planned trial.” The FDA states that “[t]he choice of M2, [the non-inferiority margin], is a matter of clinical judgement,” but specific guidance is absent. Jones et al⁴¹ proposed that at least 50% of the effect of the control regimen be retained by the experimental regimen, and D'Agostino et al³⁹ reported that the clinical community is leaning toward this percentage. D'Agostino et al³⁹ also observed that with respect to the role of clinical judgement in determining the NIM, “we have never seen a case where this has actually been employed.” In 2017 Mauri et al⁴² called for “explicit justification of the acceptable margin that is based on a measured or anticipated

benefit of the experimental treatment” and further suggested “incorporating decision analysis (population or policy perspective) or patient questionnaires designed to consider how the non-inferiority margin and expected benefit are balanced.”

An elegant feature of this work is that it provides a unified statistical framework that links the inherent risk-benefit trade-off in the DOOR-type outcome approach to the allowable trade-off reflected by the non-inferiority margin in a non-inferiority comparison approach. This established link has a consequential practical benefit by the introduction of the concept of the “maximum allowable decrease in efficacy” or MADE, a quantification of the anticipated benefit which can readily be determined under pre-specified assumptions regarding the extent of anticipated safety benefit ($-\delta_S$) of the experimental regimen over the active control and the dependence between efficacy and safety (R).

As we have shown in this work, independence between efficacy and safety leads to closed form solutions and a clearer and simpler picture of the trade-off between efficacy and safety, as well as a clear algebraic understanding of the nature of dependence of MADE on the safety parameters. However, we have also seen from results presented the existence of MADE, and that its properties and the inherent trade-off are all preserved even when there is dependence. From examples, we have shown the behavior of MADE under the full range of possible values of the dependence parameter ($0 < R < \infty$). Additionally, the simulations evaluated the operating characteristics under the full range of possible values of the dependence parameter ($0 < R < \infty$). In summary, the methods do not require an independence assumption.

In a non-inferiority trial design, using the MADE as the NIM can be used to justify the point that balances risks and benefits. However, it may also be necessary to ensure that an adequate proportion of the effect of the control regimen is maintained, for example $[P_{TE} - P_{CE}] - f \times (P_{CE} - P_{PE})$, where P_{PE} is the proportion of participants who have a favorable outcome on placebo or who remain untreated and f is the minimum proportion of the putative placebo vs control comparison that should be retained. For example, choosing $f = 0.5$ assures that at least 50% of the effect of the control regimen is preserved, but the factor can also be smaller as long as it is non-zero. The NIM can be set to the smaller of $f \times (P_{CE} - P_{PE})$ and the MADE. If conducting a superiority trial with the risk-benefit outcome approach, it will be important during the analysis to show that P_{TE} preserves the desired minimum effect, otherwise we risk having demonstrated superiority of an ineffective treatment. To do this, an assumption on favorable outcome percentage for persons receiving placebo or remain untreated need to be specified during the design stage, ideally based on data from a study in a similar population.

Non-inferiority trials must also show assay sensitivity, which is the “ability to distinguish an effective treatment from a less effective or ineffective treatment”¹⁵. This requires that “similarly designed trials in the past regularly distinguished effective treatment from less effective or ineffective treatments”¹⁵. The addition of the methodology proposed here to the toolbox brings increased statistical rigor in the design of a non-inferiority trial, and does not preclude this critical attribute.

In any non-inferiority trial, the NIM is a reflection of an “anticipated” or “perceived” benefit. This benefit remains “perceived” until data are collected from the trial itself that may corroborate or validate the extent of the benefit, if any. With the availability of the MADE as a quantifiable framework to inform the NIM, the trialist now has a way to assess how accurately the NIM reflects this benefit by checking its pre-determined value against the corresponding MADE estimate derived from the observed safety data, as well as the observed correlation between efficacy and safety, obtained at the end of the trial. As has traditionally been done, eliciting quantities from expert opinion may be quite useful in determination of the NIM, but this time they would be guided by the MADE as an available statistical framework. More useful sources such as data from early phase studies of the experimental regimen of interest, including preferences elicited from patients or their advocates, in conjunction with available data on the active control obtained from completed trials or relevant observational studies, can also be used to inform these assumptions and a NIM could be chosen that is robust to a range of plausible values of MADE. The range could potentially be quite wide, resulting in a very conservative choice. A Bayesian adaptive approach could be developed that allows an update to the NIM (as informed by MADE) during the conduct of the study, and would be an interesting direction for future research. For example, one can specify prior distributions on control and treatment safety rates, control arm efficacy rate, and their correlation, and compute the MADE averaged over the joint prior distribution. This may give a more robust value for the MADE at the design stage and, at different stages of the trial, can be directly compared with a MADE averaged over a joint posterior distribution of those values.

Whether one chooses to use the MADE to inform the NIM, or develops an alternative methodology to quantify the anticipated benefit, one cannot deny the challenges that comes with the use of a non-inferiority study design brought about by a potential discrepancy between a pre-determined NIM and the actual benefit that it is intended to reflect. In this regard, using the DOOR approach or the weighted outcome approach is quite appealing since the trade-off is inherent and already captured in the definition of the risk-benefit outcome.

The simulation study suggests modestly to significantly higher power with the conduct of a non-inferiority trial based on MADE, compared to a superiority trial based on the corresponding risk-benefit approach; if feasibility is an issue, whether in terms of ability to enroll sufficient numbers of participants or due to limited resources, then there may be an advantage with choosing a non-inferiority design, depending on the assumptions on the design parameters.

The risk-benefit approach compares the totality of a patient’s experience with a new treatment regimen vs an active control. If a risk-benefit composite outcome is the primary outcome, it remains important that disaggregated analyses of each domain in the composite outcome be reported as key secondary analyses in order to have a complete picture of the results of the trial. On the other hand, if a trial is designed primarily as a non-inferiority study, the MADE can be used to inform the setting of NIM that balances risks and benefits, and the superiority comparison using the risk-benefit outcome may serve as an important complementary analysis.

This paper focused on developing the MADE for the situation when the experimental regimen has anticipated benefits over the control in a single dimension. The concept of the MADE should be extended to the situation where an experimental regimen has gains in two or more domains over the control regimen, such as safety, tolerability, acceptability and quality of life, among others, instead of just one. We focused on the MADE for cases when the safety domain is binary; however, in reality, types of adverse events can differ in seriousness and, for a given type of event, there may be varying levels of severity. For example, different types of adverse events can be categorized into degrees of severity using the Division of AIDS (DAIDS) toxicity tables⁴³. The construction of MADE can be extended to the case when some of the domains have more than two levels, such as grading of severity in the case of safety, and corresponding properties derived accordingly and would be an interesting area for further research. Finally, the outcome of death, which has traditionally either been considered as treatment failure or a safety event (depending on the cause of death), or analyzed as a competing risk, needs consideration in the context of these approaches, especially when its occurrence is non-negligible in the study population of interest.

Acknowledgement

We would like to thank Michael Hughes for reviewing and providing helpful suggestions that led to improvements in the manuscript.

8. Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number UM1 AI068616-12 (Grace Montepiedra, Soyeon Kim) and UM1 AI068634, UM1 AI068636 and UM1 AI106701 (Ritesh Ramchandani, Sachiko Miyahara).

7.: Appendix

To prove that (1) is equivalent to (8) for the case when $R = 1$, substitute the probabilities in (7) into expressions (2) and (3) and, after some algebraic manipulations, the following expressions can be derived:

$$\begin{aligned} P(C>T) &= P_{CS} \times (1 - P_{TS}) \times [P_{CE} \times P_{TE} + (1 - P_{CE}) \times (1 - P_{TE})] + P_{CE} \\ &\times (1 - P_{TE}), \\ P(T>C) &= P_{TS} \times (1 - P_{CS}) \times [P_{CE} \times P_{TE} + (1 - P_{CE}) \times (1 - P_{TE})] + P_{TE} \\ &\times (1 - P_{CE}). \end{aligned} \quad (17)$$

Substituting (17) into (1) reduces to the following inequality:

$$(P_{CS} - P_{TS}) \times [P_{CE} \times P_{TE} + (1 - P_{CE}) \times (1 - P_{TE})] + (P_{CE} - P_{TE}) > 0,$$

which is equivalent to (8) with $w_E = P_{CE} \times P_{TE} + [(1 - P_{CE}) \times (1 - P_{TE})]$.

In order to prove Lemma 3, we need to first show $\pi_T(4)$ is monotone increasing in P_{TE} which is shown in Lemma 1 and that $P_{TE} - \pi_T(4)$ is monotone increasing in P_{TE} which is shown in Lemma 2.

Lemma 1:

Let T be the random variable defined by the DOOR ranking in Table 1 for Arm T. Suppose that T has a multinomial distribution denoted by $\{\pi_T(j), j=1,2,3,4\}$. For treatment T, let P_{TE} and P_{TS} be the marginal probability of a favorable efficacy outcome and the probability of experiencing no treatment-limiting adverse event, respectively, and let R_T be the dependence parameter between these two safety and efficacy outcomes. Let the values of P_{TS} and R_T be fixed. Then $\pi_T(4)$ is monotone increasing in P_{TE} .

Proof: If $R_T = 1$, then $\pi_T(4) = P_{TS} \times P_{TE}$ which is increasing in P_{TE} since P_{TS} is always positive. From Sozu¹⁵, if $R_T \neq 1$, $\pi_T(4)$ can be written as a function of P_{TE} as follows:

$$\pi_T(4) = [1 + (P_{TE} + P_{TS}) \times (R_T - 1) - S_T] / [2 \times (R_T - 1)],$$

where $S_T = \sqrt{[1 + (P_{TE} + P_{TS}) \times (R_T - 1)]^2 + 4 \times R_T \times (1 - R_T) \times P_{TE} \times P_{TS}}$. Taking the derivative of $\pi_T(4)$ with respect to P_{TE} we get

$$\frac{\partial \pi_T(4)}{\partial P_{TE}} = \frac{1}{2} \times \left[1 - S_T^{-1} \times \{1 + P_{TE} \times (R_T - 1) - P_{TS} \times (R_T + 1)\} \right]. \quad (18)$$

To prove the lemma, we need to show that expression (18) is greater than or equal to 0, which is true whenever expression

$$[1 + P_{TE} \times (R_T - 1) - P_{TS} \times (R_T + 1)] \quad (19)$$

is less than or equal to 0. Expression (18) is greater than 0 if and only if expression (19) is less than S_T or, equivalently, if and only if

$$4 \times P_{TS} \times R_T \times [2 - P_{TE} - P_{TS} + P_{TE} \times (R_T - 1)] > 0,$$

which is true whenever expression (19) is greater than 0, since $2 - P_{TE} - P_{TS} + P_{TE} \times (R_T - 1)$ is greater than (19).

Lemma 2:

Let T be the random variable defined by the DOOR ranking in Table 1 for Arm T. Suppose that T has a multinomial distribution denoted by $\{\pi_T(j), j=1,2,3,4\}$. For treatment T, let P_{TE} and P_{TS} be the marginal probability of a favorable efficacy outcome and the probability of experiencing no treatment-limiting adverse event, respectively, and let R_T be the dependence parameter between these two safety and efficacy outcomes. Let the values of P_{TS} and R_T be fixed. Then $P_{TE} - \pi_T(4)$ is monotone increasing in P_{TE} .

Proof: If $R_T = 1$, then $P_{TE} - \pi_T(4) = (1 - P_{TS}) \times P_{TE}$ which is increasing in P_{TE} since $1 - P_{TS}$ is always positive. If $R_T \neq 1$, then the derivative of $P_{TE} - \pi_T(4)$ with respect to P_{TE} is 1 minus expression (18), which is

$$\frac{1}{2} + \frac{1}{2} \times \left[S_T^{-1} \times \{1 + P_{TE} \times (R_T - 1) - P_{TS} \times (R_T + 1)\} \right]. \quad (20)$$

Using a similar argument as in the proof for Lemma 1, it can be shown that expression (20) is greater than or equal to zero.

Lemma 3:

Let C and T be random variables defined by the DOOR rankings in Table 1 for Arms C and T, respectively. Suppose that C and T have multinomial distributions denoted by $\{\pi_C(i), i=1,2,3,4\}$ and $\{\pi_T(j), j=1,2,3,4\}$, respectively. Let P_{CE} , P_{CS} , P_{TE} and P_{TS} be defined as in (4) and (5) and fall in the range (0, 1). Let the distribution of C be fixed and $\pi_C(i) > 0$ for all i. Furthermore, let $P_{TS} > 0$, the marginal safety probability for T, and the dependence parameter between safety and efficacy, denoted by R_T , be fixed quantities. Then $P(T > C) - P(C > T)$ is monotone increasing in P_{TE} .

Proof: Substitute the following expressions into (2) and (3), and then into (1):

$$\begin{aligned} \pi_C(1) &= 1 - P_{CS} - P_{CE} + \pi_C(4), \\ \pi_C(2) &= P_{CS} - \pi_C(4), \\ \pi_C(3) &= P_{CE} - \pi_C(4), \end{aligned}$$

$$\begin{aligned} \pi_T(1) &= 1 - P_{TS} - P_{TE} + \pi_T(4), \\ \pi_T(2) &= P_{TS} - \pi_T(4), \\ \pi_T(3) &= P_{TE} - \pi_T(4), \end{aligned}$$

to obtain the following:

$$\begin{aligned} f[P_{TE}] = P(T > C) - P(C > T) &= [P_{CE} - \pi_C(4)] \times P_{TE} + [(P_{CS} - \pi_C(4)) + P_{CE}] \times \pi_T(4) + [(1 - P_{CE}) \\ &+ [(P_{CS} - \pi_C(4))] \times [P_{TE} - \pi_T(4)]]. \end{aligned}$$

We need to show that $f[P_{TE}]$ is monotone increasing in P_{TE} . The first term, which is $[P_{CE} - \pi_C(4)] \times P_{TE}$, is monotone increasing in P_{TE} since $P_{CE} - \pi_C(4)$ is at least 0. The second term, which is $[(P_{CS} - \pi_C(4)) + P_{CE}] \times \pi_T(4)$ is monotone increasing in P_{TE} since $P_{CS} - \pi_C(4)$ and P_{CE} are always at least 0 and, from Lemma 1, $\pi_T(4)$ is monotone increasing in P_{TE} . Finally, the third term, which is $[(P_{CS} - \pi_C(4))] \times [P_{TE} - \pi_T(4)]$, is monotone increasing in P_{TE} since $P_{CS} - \pi_C(4)$ is always greater than or equal to 0 and, from Lemma 2, $P_{TE} - \pi_T(4)$ is monotone increasing in P_{TE} .

Lemma 4:

Let C_W and T_W be random variables defined by the weighted composite outcomes for Arm C and Arm T, respectively, corresponding to the DOOR outcomes defined in Table 1. Suppose that C_W and T_W have multinomial distributions denoted by (13). Let P_{CE} , P_{CS} , P_{TE}

and P_{TS} be defined as in (4), (5) and (6). Let the distribution of C be fixed and $\pi_C(i) > 0$ for all i . Furthermore, let $P_{TS} > 0$, the marginal safety probability for T , and the dependence parameter between safety and efficacy, denoted by R_T , be fixed quantities. Then $\mu_Y - \mu_X$ is monotone increasing in P_{TE} .

Proof: Substitute the expressions for $\pi_C(1)$, $\pi_C(2)$, $\pi_C(3)$, $\pi_C(4)$, $\pi_T(1)$, $\pi_T(2)$ and $\pi_T(3)$ provided in the proof of Lemma 1 into (14), and then into (15) to obtain the following:

$$f_2[P_{TE}] = \mu_T - \mu_C = (w_3 - w_1) \times [P_{TE} - P_{CE}] + (w_1 - w_2 - w_3 + w_4) \times \pi_T(4) + h$$

where $h = (w_2 - w_1) \times [P_{TS} - P_{CS}] - (w_1 - w_2 - w_3 - w_4) \times \pi_C(4)$ does not contain terms with P_{TE} . This function can be re-written as

$$f_2[P_{TE}] = (w_3 - w_1) \times [P_{TE} - \pi_T(4)] + (w_4 - w_2) \times \pi_T(4) + h.$$

From Lemmas 1 and 2, $P_{TE} - \pi_T(4)$ and $\pi_T(4)$ are both monotone increasing in P_{TE} in the range $(0, 1)$, and since both $w_3 - w_1$ and $w_4 - w_2$ are by definition greater than or equal to zero, then $f_2[P_{TE}]$ is monotone increasing in P_{TE} .

10. References

1. Madhi SA, Nachman S, Violari A, Kim S, Cotton MF, Bobat R, et al. Primary Isoniazid Prophylaxis against Tuberculosis in HIV-Exposed Children. *N Engl J Med* 2011;365:21–31. [PubMed: 21732834]
2. Martinson NA, Barnes GL, Moulton LH, Msandiwa R, Hausler H, Ram M, et al. New Regimens to Prevent Tuberculosis in Adults with HIV Infection. *N Engl J Med* 2011;365: 11–20. [PubMed: 21732833]
3. Sterling TR, Villarino E, Borisov AS, Shang N, Gordin F, Bliven-Sizemore E, et al. Three Months of Rifapentine and Isoniazid for Latent Tuberculosis Infection. *N Engl J Med* 2011;365:2155–66. [PubMed: 22150035]
4. Churchyard GJ, Fielding KL, Lewis JJ, Coetzee L, Corbett EL, Godfrey-Faussett P, et al. A Trial of Mass Isoniazid Preventive Therapy for Tuberculosis Control. *N Engl J Med* 2014; 370:301–310. [PubMed: 24450889]
5. Merle CS, Fielding K, Sow OB, Gninafon M, Lo MB, Mthiyane T, et al. A Four-Month Gatifloxacin-Containing Regimen for Treating Tuberculosis. *N Engl J Med* 2014;371(17): 1588–98. [PubMed: 25337748]
6. Jindani A, Harrison TS, Nunn AJ, Phillips PP, Churchyard GJ, Charalambous S, et al. High-dose rifapentine with moxifloxacin for pulmonary tuberculosis. *N Engl J Med* 2014; 371(17):1599–608. [PubMed: 25337749]
7. Gillespie SH, Crook AM, McHigh TD, Mendel CM, Meredith SK, Murray SR, et al. Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis. *N Engl J Med* 2014;371(17):1577–87. [PubMed: 25196020]
8. Menzies D, Adjobimey M, Ruslami R, Trajman A, Sow O, Kim H, et al. Four Months of Rifampin or Nine Months of Isoniazid for Latent Tuberculosis in Adults. *N Engl J Med* 2018;379:440–53. [PubMed: 30067931]
9. Diallo T, Adjobimey M, Ruslami R, Trajman A, Sow O, Baah JO, et al. Safety and Side Effects of Rifampin versus Isoniazid in Children. *N Engl J Med* 2018;379:454–463. [PubMed: 30067928]
10. Swindells S, Ramchandani R, Gupta A, Benson CA, Leon-Cruz J, Mwelase N, et al. One Month of Rifapentine plus Isoniazid to Prevent HIV-Related Tuberculosis. *N Engl J Med* 2019;380:1001–11. [PubMed: 30865794]

11. Nunn AJ, Phillips PPJ, Meredith SK, Chiang C-Y, Conradie F, Dalai D, et al. A Trial of a Shorter Regimen for Rifampin-Resistant Tuberculosis. *N Engl J Med* 2019;380:1201–13. [PubMed: 30865791]
12. Gupta A, Montepiedra G, Aaron L, Theron G, McCarthy K, Bradford S, et al. Isoniazid Preventive Therapy in HIV-Infected Pregnant and Postpartum Women. *N Engl J Med* 2019;381:1333–46. [PubMed: 31577875]
13. Phillips PJ, Mitnick CD, Neaton JD, Nahid P, Lienhardt C, Nunn AJ. Keeping phase III tuberculosis trials relevant: adapting to a rapidly changing landscape. *PLoS Med* 2019;16(3):e1002767 10.1371/journal.pmed.1002767 [PubMed: 30901331]
14. Olliaro PL, Vaillant M. Designing noninferiority tuberculosis treatment trials: Identifying practical advantages for drug regimens with acceptable effectiveness. *PLoS Medicine* 2019;16(7):e1002850 10.1371/journal.pmed.1002850 [PubMed: 31299047]
15. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). 11 2016 Clinical/Medical <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm>
16. Chuang-Stein C, Mohberg NR. Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Stat Med* 1991;10:1349–1359. [PubMed: 1925166]
17. Buyse M Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* 2010;29:3245–57. [PubMed: 21170918]
18. Pocock SJ, Ariti C, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 2012;33:176–82. [PubMed: 21900289]
19. Evans SR, Rubin D, Follmann D, Pennello G, Huskins WC, Powers JH, Schoenfeld D, et al. Desirability of Outcome Ranking (DOOR) and Response Adjusted for Duration of Antibiotic Risk (RADAR). *Clin Infect Dis* 2015;61(5):800–6. [PubMed: 26113652]
20. Hallstrom PH, Litwin PE, Weaver WD. A method of assigning scores to the components of a composite outcome: an example from the MITI trial. *Control Clin Trials* 1992;13:148–155. [PubMed: 1316829]
21. Bakal JA, Westerhout CM, Armstrong PW. Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled clinical trials. *Stat Methods Med Res* 2015;24(6):980–988. [PubMed: 22275378]
22. Evans SR, Follmann D. Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit:risk evaluation. *Stat Biopharm Res* 2016;8(4):386–393. [PubMed: 28435515]
23. Duc AN, Wolbers M. Weighted analysis of composite endpoints with simultaneous inference for flexible weight constraints. *Stat Med* 2017;36:442–54. [PubMed: 27782312]
24. Miyahara S, Ramchandani R, Kim S, Evans SR, Gupta A, Swindells S, et al. Applying risk-benefit analysis to outcomes in TB clinical trials. *Clin Infect Dis* 2020;70(4):698–703. [PubMed: 31414121]
25. Montepiedra G, Yuen CM, Rich ML, Evans SR. Totality of outcomes: A different paradigm in assessing interventions for treatment of tuberculosis. *J Clin Tuberc Other Mycobact Dis* 2016;4:9–13. [PubMed: 28042610]
26. Fligner MA, Policello II GE. Robust rank procedures for the Behrens-Fisher problem. *J Am Stat Assoc* 1981;76:162–168.
27. Brunner E, Neumann N. Two-sample rank tests in general linear models. *Biom J* 1986;28:395–402.
28. Huang P, Woolson RF, O'Brien PC. A rank-based sample size method for multiple outcomes in clinical trials. *Stat Med* 2008;27(16):3084–3104. [PubMed: 18189338]
29. Ramchandani R, Schoenfeld DA, Finkelstein DM. Global rank tests for multiple, possibly censored, outcomes. *Biometrics* 2016;72(3):926–935. [PubMed: 26812695]
30. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. *Stat Med* 2010;29:2169–79. [PubMed: 20687162]

31. Rauch G, Jahn-Eimermacher A, Brannath W, Kieser M. Opportunities and challenges of combined effect measures based on prioritized outcomes. *Stat Med* 2012;33:1104–1120.
32. Seddon JA, Weld ED, Schaaf HS, Garcia-Prats AJ, Kim S, Hesselning AC. Conducting efficacy trials in children with MD-TB: what is the rationale and how should they be done? *Int J Tuberc Lung Dis* 2018;22(5):24–33. *Pediatric Landscape Supplement* 10.5588/ijtld.17.0359 [PubMed: 29665950]
33. Kim S, Seddon JA, Garcia-Prats AJ, Montepiedra G. Statistical considerations for pediatric multidrug-resistant tuberculosis efficacy trials. *Int J Tuberc Lung Dis* 2018;22(5):34–39. *Pediatric Landscape Supplement* 10.5588/ijtld.17.0358
34. World Health Organization. Treatment guidelines for drug-resistant tuberculosis 2016 update. WHO/HTM/TB/2016.04. Geneva, Switzerland: WHO, 2016.
35. Aung KJ, Van Deun A, Declercq E, Sarker MR, Das PK, Hossain MA, et al. Successful ‘9-month Bangladesh regimen’ for multidrug-resistant tuberculosis among over 500 consecutive patients. *Int J Tuberc Lung Dis*. 2014;18(10):1180–7. [PubMed: 25216831]
36. Van Deun A, Maug AK, Salim MA, Das PK, Sarker MR, Daru P, et al. Short, highly effective, and inexpensive standardized treatment of multidrug-resistant tuberculosis. *Am J Respir Crit Care Med*. 2010;182(5):684–92. [PubMed: 20442432]
37. Seddon JA, Hesselning AC, Godfrey-Faussett P, Schaaf HS. High treatment success in children treated for multidrug-resistant tuberculosis: an observational cohort study. *Thorax*. 2014;69(5):458–64. [PubMed: 24064441]
38. Bakayoko A, Mbulula M-Lo, Noeske Jr, Ciza Fo, Souleymane B, Fikouma V, et al. Treatment of multidrug-resistant tuberculosis in children and adolescents with a 9-month regimen in Africa. Durban, South Africa: TB2016 Conference; 2016 [Available from: <http://www.tb2016.org/Portals/2/File/Posters/O12.pdf?ver=2016-08-18-165859-717>.]
39. D’Agostino SR, RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Stat Med* 2003;22:169–186. [PubMed: 12520555]
40. European Medicines Agency Committee for Medicinal Products for Human Use (CHMP) Guideline on the Choice of the Non-inferiority Margin. London, 27 July 2005 Doc. Ref. EMEA/CPMP/EWP/2158/99.
41. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996;313:36–39. [PubMed: 8664772]
42. Mauri L, D’Agostino Sr RB. Challenges in the Design and Interpretation of Noninferiority Trial. *N Engl J Med* 2017;377:1357–67. [PubMed: 28976859]
43. U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Allergy and Infectious Diseases, Division of AIDS. Division of AIDS (DAIDS) Table for Grading the Severity of Adult and Pediatric Adverse Events, Corrected Version 2.1. [7 2017]. Available from: <https://rsc.niaid.nih.gov/sites/default/files/daidsgradingcorrectedv21.pdf>

Figure 1:
Impact of dependence on safety and efficacy on effect size (%).

Figure 2:
Maximum Allowable Decrease in Efficacy by dependence on safety and efficacy and expected improvement in safety.

Table 1:

A DOOR ranking of the composite efficacy / safety outcome

Ranking (4 = Best, 1 = Worst)	Composite Outcome		
	Efficacy		Safety
1	Unfavorable outcome	and	Treatment-limiting adverse event
2	Unfavorable outcome	and	No treatment-limiting adverse event
3	Favorable outcome	and	Treatment-limiting adverse event
4	Favorable outcome	and	No treatment-limiting adverse event

Table 2a:

Example of joint probabilities of the DOOR outcome for each treatment arm

DOOR outcome, j	Probabilities	
	Arm C, $\pi_C(j)$	Arm T, $\pi_T(j)$
1	3.00%	0.65%
2	12.00%	12.35%
3	17.00%	4.35%
4	68.00%	82.65%

Table 2b:

Marginal probabilities and odds ratio corresponding to the joint probabilities in Table 2a assuming independence.

Parameter	Treatment C	Treatment T
Probability of favorable outcome	85%	87%
Probability of no treatment-limiting AE	80%	95%
Odds ratio (R)	1.0	1.0

Table 3.

Trade-off between efficacy and safety when $R_C = R_T = 1$ (efficacy and safety are independent outcomes in both arms) and $P_{CE} = 85\%$

$-\delta_S = P_{TS} - P_{CS}$	P_{TE}	$\delta_E = P_{CE} - P_{TE}$	w	$-\delta_S \times w_E$	Effect size	$\delta_E^* = \text{MADE}$
5%	87%	-2%	<i>0.759</i>	<i>3.80%</i>	<i>5.80%</i>	3.60%
	86%	-1%	<i>0.752</i>	<i>3.76%</i>	<i>4.76%</i>	
	85%	0%	<i>0.745</i>	<i>3.73%</i>	<i>3.73%</i>	
	84%	1%	<i>0.738</i>	<i>3.69%</i>	<i>2.69%</i>	
	83%	2%	<i>0.731</i>	<i>3.66%</i>	<i>1.66%</i>	
	82%	3%	<i>0.724</i>	<i>3.62%</i>	<i>0.62%</i>	
	$P_{TE}^* = 81.4\%$	3.60%	0.719	3.60%	0.00%	
	81%	4%	0.717	3.59%	-0.41%	
	80%	5%	0.710	3.55%	-1.45%	
10%	87%	-2%	<i>0.759</i>	<i>7.59%</i>	<i>9.59%</i>	6.97%
	85%	0%	<i>0.745</i>	<i>7.45%</i>	<i>7.45%</i>	
	83%	2%	<i>0.731</i>	<i>7.31%</i>	<i>5.31%</i>	
	81%	4%	<i>0.717</i>	<i>7.17%</i>	<i>3.17%</i>	
	79%	6%	<i>0.703</i>	<i>7.03%</i>	<i>1.03%</i>	
	$P_{TE}^* = 78.1\%$	6.97%	0.697	6.97%	0.00%	
	77%	8%	0.689	6.89%	-1.11%	
	75%	10%	0.675	6.75%	-3.25%	
15%	87%	-2%	<i>0.759</i>	<i>11.39%</i>	<i>13.39%</i>	10.12%
	85%	0%	<i>0.745</i>	<i>11.18%</i>	<i>11.18%</i>	
	82%	3%	<i>0.724</i>	<i>10.86%</i>	<i>7.86%</i>	
	79%	6%	<i>0.703</i>	<i>10.55%</i>	<i>4.55%</i>	
	76%	9%	<i>0.682</i>	<i>10.23%</i>	<i>1.23%</i>	
	$P_{TE}^* = 74.9\%$	10.12%	0.674	10.12%	0.00%	
	73%	12%	0.661	9.92%	-2.08%	
	70%	15%	0.640	9.60%	-5.40%	

Entries in *italics* are scenarios when treatment T is superior to treatment C

Table 4.

MADE under varying odds ratios comparing safety events and efficacy ($R_C = R_T = R$) and safety risk difference [i.e., $-\delta_S = P_{TS} - P_{CS}$] assuming control arm rates $P_{CE} = 85\%$, $P_{CS} = 80\%$.

$-\delta_S = P_{TS} - P_{CS}$	Odds ratio (R)	P^*_{TE}	$\delta_E^* = \text{MADE}$
15%	∞	78.6%	6.4%
	10.0	76.2%	8.8%
	5.0	75.7%	9.3%
	2.0	75.2%	9.8%
	1.0	74.9%	10.1%
	0.5	74.7%	10.3%
	0.2	74.5%	10.5%
	0.1	74.4%	10.6%
	0	74.4%	10.6%
10%	∞	79.4%	5.6%
	10.0	78.2%	6.8%
	5.0	78.2%	6.8%
	2.0	78.1%	6.9%
	1.0	78.0%	7.0%
	0.5	78.0%	7.0%
	0.2	78.0%	7.0%
	0.1	77.9%	7.1%
	0	77.9%	7.1%
5%	∞	80.3%	4.7%
	10.0	81.0%	4.0%
	5.0	81.2%	3.8%
	2.0	81.3%	3.7%
	1.0	81.4%	3.6%
	0.5	81.4%	3.6%
	0.2	81.4%	3.6%
	0.1	81.5%	3.5%
	0	81.5%	3.5%

Table 5.

MADE using the weighted component outcome

ω_E	$k(\omega_E)$	$\delta_S = P_{TS} - P_{CS}$	MADE
0.5	1.0	5%	5.00%
		10%	10.00%
		15%	15.00%
0.6	0.67	5%	3.35%
		10%	6.70%
		15%	10.05%
0.7	0.43	5%	2.15%
		10%	4.30%
		15%	6.45%
0.8	0.25	5%	1.25%
		10%	2.50%
		15%	3.75%
0.9	0.11	5%	5.50%
		10%	1.10%
		15%	1.65%

Table 6.1:

Simulations to estimate power to conclude (i) experimental arm superiority with respect to DOOR outcome versus (ii) non-inferiority on efficacy, when $P_{TE} = 0.83$ (slightly worse than $P_{CE} = 0.85$) and experimental arm has better safety rate ($-\delta_S > 0$).

$-\delta_S =$ $P_{TS} -$ P_{CS}	n per arm	R	$P(T > C) -$ $P(C > T)$	MADE	P^*_{TE}	Power (%)		
						Wilcoxon rank- sum test (one- sided) with ranked outcome	Non-inferiority on efficacy using NIM=MADE	Superiority on safety and non-inferiority on efficacy using NIM=MADE
0.05	1300	0	0.018	0.035	0.815	16.6	17.6	15.3
		0.1	0.018	0.035	0.815	16.1	17.2	15.0
		0.2	0.018	0.036	0.814	15.4	17.8	15.5
		0.5	0.017	0.036	0.814	14.8	18.0	16.0
		1	0.016	0.036	0.814	13.8	17.6	16.0
		2	0.016	0.037	0.813	14.0	19.4	18.1
		5	0.016	0.038	0.812	13.8	22.2	21.4
		10	0.016	0.04	0.81	14.9	25.9	25.2
		∞	0.024	0.047	0.803	37.2	46.0	46.0
0.10	400	0	0.061	0.071	0.779	44.8	45.9	44.3
		0.1	0.06	0.071	0.779	43.6	46.5	44.8
		0.2	0.059	0.07	0.78	43.1	46.3	44.7
		0.5	0.056	0.07	0.78	39.8	44.3	43.0
		1	0.053	0.07	0.78	37.1	44.8	43.6
		2	0.05	0.069	0.781	34.4	44.0	43.2
		5	0.046	0.068	0.782	31.4	42.8	42.3
		10	0.043	0.068	0.782	28.7	41.8	41.5
		∞	0.032	0.056	0.794	21.2	25.1	25.1
0.15	200	0	0.104	0.106	0.744	62.7	61.0	60.8
		0.1	0.101	0.106	0.744	60.2	60.2	59.9
		0.2	0.1	0.105	0.745	58.9	58.2	58.0
		0.5	0.095	0.103	0.747	54.9	57.0	56.7
		1	0.09	0.101	0.749	52.6	55.4	55.2
		2	0.083	0.098	0.752	45.7	51.2	51.0
		5	0.073	0.093	0.757	41.3	47.9	47.9
		10	0.066	0.088	0.762	35.2	40.8	40.7
		∞	0.04	0.064	0.786	17.6	19.2	19.2

Table 6.2:

Simulations to estimate power to conclude (i) experimental arm superiority with respect to DOOR outcome versus (ii) non-inferiority on efficacy, when $P_{TE} = 0.87$ (slightly better than $P_{CE} = 0.85$) and experimental arm has better safety rate ($-\delta_S > 0$).

$-\delta_S = P_{TS} - P_{CS}$	n per arm	R	$P(T>C) - P(C>T)$	MADE	P^*_{TE}	Power (%)		
						Wilcoxon rank-sum test (one-sided) with ranked outcome	Non-inferiority on efficacy using NIM=MADE	Superiority on safety and Non-inferiority on efficacy using NIM=MADE
0.05	1300	0	0.066	0.035	0.815	94.5	97.8	88.6
		0.1	0.065	0.035	0.815	94.1	98.0	89.2
		0.2	0.064	0.036	0.814	93.8	98.1	89.1
		0.5	0.061	0.036	0.814	91.8	98.0	89.7
		1	0.058	0.036	0.814	89.9	98.3	89.7
		2	0.054	0.037	0.813	86.1	98.4	89.8
		5	0.05	0.038	0.812	81.5	98.6	90.0
		10	0.047	0.04	0.81	79.3	99.2	91.0
		∞	0.046	0.047	0.803	88.4	99.8	91.8
0.10	400	0	0.109	0.071	0.779	91.8	94.8	92.3
		0.1	0.107	0.071	0.779	91.1	94.8	92.1
		0.2	0.105	0.07	0.78	90.4	95.1	92.6
		0.5	0.101	0.07	0.78	88.2	94.4	91.9
		1	0.096	0.07	0.78	85.7	94.5	92.1
		2	0.09	0.069	0.781	83.2	94.6	92.5
		5	0.082	0.068	0.782	77.2	94.4	92.3
		10	0.077	0.068	0.782	74.3	93.5	91.5
		∞	0.068	0.056	0.794	73.2	84.5	84.4
0.15	200	0	0.152	0.106	0.744	92.7	93.8	93.5
		0.1	0.149	0.106	0.744	92.2	93.5	93.2
		0.2	0.146	0.105	0.745	91.5	92.9	92.6
		0.5	0.141	0.103	0.747	89.6	93.0	92.5
		1	0.134	0.101	0.749	86.9	91.8	91.4
		2	0.125	0.098	0.752	83.3	90.1	89.7
		5	0.113	0.093	0.757	78.5	88.4	88.2
		10	0.104	0.088	0.762	72.2	84.1	83.9
		∞	0.076	0.064	0.786	53.5	63.0	63.0

Table 6.3:

Simulations to estimate power to conclude (i) experimental arm superiority with respect to DOOR outcome versus (ii) non-inferiority on efficacy, when $P_{TE} = 0.85$ (equal to $P_{CE} = 0.85$) and experimental arm has better safety rate ($-\delta_S > 0$).

$-\delta_S = P_{TS} - P_{CS}$	n per arm	R	$P(T>C) - P(C>T)$	MADE	P^*_{TE}	Power (%)		
						Wilcoxon rank sum test (one-sided) with ranked outcome	Non-inferiority on efficacy using NIM=MADE	Superiority on safety and non-inferiority on efficacy using NIM=MADE
0.05	1300	0	0.042	0.035	0.815	62.3	68.7	61.8
		0.1	0.042	0.035	0.815	60.2	69.3	62.3
		0.2	0.041	0.036	0.814	59.2	70.7	63.7
		0.5	0.039	0.036	0.814	57.1	70.6	64.2
		1	0.037	0.036	0.814	51.9	70.7	64.5
		2	0.035	0.037	0.813	41.9	72.5	66.6
		5	0.033	0.038	0.812	46.7	76.0	70.7
		10	0.032	0.04	0.81	45.1	79.1	74.3
		∞	0.042	0.047	0.803	82.0	91.8	89.4
0.1	400	0	0.085	0.071	0.779	73.3	76.7	74.5
		0.1	0.083	0.071	0.779	71.8	76.6	74.2
		0.2	0.082	0.07	0.78	70.7	77.6	75.4
		0.5	0.078	0.07	0.78	67.5	76.0	73.9
		1	0.074	0.07	0.78	64.1	76.3	74.3
		2	0.07	0.069	0.781	60.4	75.9	74.2
		5	0.064	0.068	0.782	54.7	74.4	73.1
		10	0.06	0.068	0.782	51.8	73.2	72.2
		∞	0.05	0.056	0.794	45.4	55.2	55.2
0.15	200	0	0.128	0.108	0.742	81.3	81.2	81.0
		0.1	0.125	0.107	0.743	79.2	80.6	80.3
		0.2	0.123	0.106	0.744	77.8	79.8	79.4
		0.5	0.118	0.104	0.746	75.4	79.0	78.6
		1	0.112	0.101	0.749	71.7	76.8	76.4
		2	0.104	0.098	0.752	66.8	74.0	73.7
		5	0.093	0.093	0.757	60.8	70.8	70.1
		10	0.085	0.09	0.76	54.1	64.2	64.1
		∞	0.058	0.072	0.778	32.9	38.4	38.4

Table 7:

Simulations to estimate type I error to test (i) experimental arm superiority with respect to DOOR outcome versus (ii) non-inferiority on efficacy, when experimental arm has better safety rate ($-\delta_S > 0$).

$-\delta_S =$ $P_{TS} - P_{CS}$	n per arm	R	$P(T>C) - P(C>T)$	MADE	P^*_{TE}	Type I error (%)		
						Wilcoxon rank-sum test	Non-inferiority on efficacy with NIM=MADE	Superiority on safety and non-inferiority on efficacy using NIM=MADE
0.05	1300	0	0	0.035	0.815	2.5	2.1	1.8
		0.1	0	0.035	0.815	2.5	2.1	1.8
		0.2	0	0.036	0.814	2.3	2.2	1.9
		0.5	0	0.036	0.814	2.5	2.3	2.1
		1	0	0.036	0.814	2.3	1.9	1.7
		2	0	0.037	0.813	2.7	2.4	2.3
		5	0	0.038	0.812	2.4	2.0	2.0
		10	0	0.04	0.81	2.7	2.2	2.2
		∞	0	0.047	0.803	2.8	2.3	2.3
0.1	400	0	0	0.071	0.779	2.6	2.1	1.9
		0.1	0	0.071	0.779	2.6	1.8	1.7
		0.2	0	0.07	0.78	2.8	2.4	2.3
		0.5	0	0.07	0.78	2.6	2.0	1.8
		1	0	0.07	0.78	2.5	2.0	2.0
		2	0	0.069	0.781	2.6	2.0	2.0
		5	0	0.068	0.782	2.9	2.1	2.1
		10	0	0.068	0.782	2.7	2.4	2.4
		∞	0	0.056	0.794	2.4	2.3	2.3
0.15	200	0	0	0.106	0.744	2.9	1.8	1.8
		0.1	0	0.106	0.744	2.8	1.9	1.8
		0.2	0	0.105	0.745	2.7	1.8	1.8
		0.5	0	0.103	0.747	2.8	2.1	2.1
		1	0	0.101	0.749	2.8	2.1	2.1
		2	0	0.098	0.752	2.7	2.0	2.0
		5	0	0.093	0.757	2.6	2.0	2.0
		10	0	0.088	0.762	2.4	1.7	1.7
		∞	0	0.064	0.786	2.5	2.0	2.0