The New England Journal of Statistics in Data Science Volume 2, 104–111 (2024)
DOI: https://doi.org/10.51387/24-NEJSDS57

Non-inferiority Clinical Trials: Treating Margin as Missing Information

Yulia SIDI, Benjamin STOCKTON, and Ofer HAREL*

Abstract

Non-inferiority (NI) clinical trials' goal is to demonstrate that a new treatment is not worse than a standard of care by a certain amount called margin. The choice of non-inferiority margin is not straightforward as it depends on historical data, and clinical experts' opinion. Knowing the "true", objective clinical margin would be helpful for design and analysis of non-inferiority trials, but it is not possible in practice. We propose to treat non-inferiority margin as missing information. In order to recover an objective margin, we believe it is essential to conduct a survey among a group of representative clinical experts. We introduce a novel framework, where data obtained from a survey are combined with NI trial data, so that both an estimated clinically acceptable margin and its uncertainty are accounted for when claiming non-inferiority. Through simulations, we compare several methods for implementing this framework. We believe the proposed framework would lead to better informed decisions regarding new potentially non-inferior treatments and could help resolve current practical issues related to the choice of the margin.

KEYWORDS AND PHRASES: Incomplete data, Margin justification, Multiple imputation, Non-inferiority, Survey.

1. INTRODUCTION

While the number of non-inferiority (NI) clinical trials continues to grow, design and analysis of such trials remains challenging. Unlike superiority trials, where the goal is to show that a new treatment is better than a control, NI trials seek to demonstrate that a new treatment is not worse than a standard therapy by an acceptable margin [10]. In order to offset such acceptable loss of standard treatment effect, a non-inferior agent is expected to offer other benefits, such as less severe adverse events, improved drug adherence and/or lower costs. NI trial design is usually considered when using placebo is unethical, as delaying treatment with standard care would cause irreversible health damage or death.

The choice of NI margin is not straightforward as it relies heavily on both historical data, and clinical experts opinion [5, 18]. As described by [10], at the first step, one needs to determine the standard treatment effect over placebo (M_1), using usually a meta-analysis of historical data. Then, a clinically acceptable margin (M_2), which has to be strictly lower than M_1 is chosen by clinical experts. A common analysis strategy for a NI trial is carried out using 95%–95% confidence interval (CI) approach. The first 95% CI corresponds to the lower/upper bound of the standard treatment effect over placebo from meta-analysis of historical trials, while the second 95% CI represents a comparison between the new non-inferior treatment and standard of care in the current NI trial [10]. The lower/upper bound of the later 95% CI is the one compared to M_2 in order to determine non-inferiority. This strategy is also called a "fixed margin" approach, due to the fact that the margin is set during the design stage, and is used for the final inference of the study.

Although the determination of the margin has been extensively discussed in the literature [17, 14, 15, 24, 16, 13, 22], the reasons for choosing a specific margin remain poorly reported in practice. According to systematic reviews of published NI and equivalence trials, margin justification was mentioned by 45.7%, 23%, 45%, 42.1% and 38% as reported by [39, 34, 28, 2, 25] respectively. These findings underline challenges associated with the choice of a margin for NI trials. Obviously, just determination of M_1 is very complex, since historical data carries publication bias and the previously observed treatment effect embeds some level of uncertainty. However, even if the standard treatment effect is maintained in the current NI study and the study has assay sensitivity, it is not clear how to choose one number M_2, so that it will be clinically acceptable. A legitimate question that arises here is the degree of subjectivity of the margin choice. Would it be sufficient to discuss the margin with only one clinical expert? What if an investigator who conducts an NI study reaches out to five clinical experts and they all provide different opinions, how should these opinions be incorporated into the current practices of design and analysis of the study?

If we can obtain opinions regarding a clinically acceptable margin from all clinical experts, these will constitute to a margin population. Within the margin population, there is a "true", objective M_2, which for example, could be set as a

*Corresponding author.

mean opinion across all clinical experts. Knowing the "true", objective M_2 would be extremely helpful for design and analysis of NI trials, however since the "true" margin cannot be observed, we propose to treat it as missing information. We believe that in order to make a proper inferences regarding non-inferiority of the new treatment compared to a standard of care, while minimizing subjectivity of the margin choice it is imperative to conduct a survey upon clinical experts in this regard. Such survey data can be used to make an informed decision regarding NI of the new treatment.

In this paper, we present a general framework for combining results from a clinical experts survey and NI study. Ideally, the clinical experts survey should consists of a representative sample of clinicians. Obviously, such assumption could be violated in practice by either surveying a very small number of clinicians, and/or by obtaining opinions of, for instance, more conservative experts. If clinicians conservatism or lack of thereof in respect to a clinical margin is related to other data for the representative sample (professional or demographic characteristics of the clinicians), such data could then be utilized to achieve an objective NI decision. In order to reach this goal, we propose to use multiple imputation (MI) approach [30, 21] within the above framework.

MI is a principled approach and is known to handle well incomplete data, a comprehensive review and general implementation of MI can be found in [33, 12, 27]. Within our framework, unobserved clinical experts opinions correspond to incomplete data, while professional or demographic characteristics correspond to the information used to impute the unobserved opinions. If clinical experts opinions are indeed related to their professional or demographic characteristics then MI is expected to produce reliable inferences about the parameter of interest, and therefore lead to an objective NI decision.

In the Section 2 we present our novel framework, along with simulation set-up for assessing performance of several methods within the framework. Section 3 shows results of the simulation, while Section 4 provides discussion and conclusions.

2. METHODS

2.1 Fraction Preservation as a Random Variable

Let $Y_{ij} \sim Bernoulli(p_i)$ be an occurrence of a favorable event (such as healing from a disease) for subject j, in a treatment group i. $j = 1 \dots N_i$, where N_i is a sample size of group i and $i = C, T$ represents control (or standard), and new treatment respectively. p_i is the true proportion of favorable events in group i. The hypothesis of interest is of the following form:

$$H_0 : p_C - p_T \geq M_2 \quad vs \quad H_1 : p_C - p_T < M_2, \tag{2.1}$$

where M_2 is a clinically acceptable margin, which usually constitutes a fraction of the previously observed control treatment effect over placebo M_1. In other words: $M_2 = (1-\lambda)M_1$, where λ is the fraction of the control treatment effect which clinical experts consider justifiable. We assume that M_1 has been determined based on historical studies and is fixed at the time the non-inferiority trial is being designed, and λ follows some distribution F with mean μ_λ and variance σ^2_λ. While for a known distribution F, any function of random variable λ can be used to construct the null and alternative hypotheses to test non-inferiority, we will focus on μ_λ throughout this article, since population mean is a commonly used parameter of interest in many practical situations. Following the notation above we can re-write the hypothesis in (2.1) as:

$$H_0 : p_C - p_T \geq (1-\mu_\lambda)M_1 \quad vs \quad H_1 : p_C - p_T < (1-\mu_\lambda)M_1. \tag{2.2}$$

For a known population distribution F, we demonstrate how the value of the margin could significantly impact study design in terms of sample size calculation. A sample size per treatment arm (n) can be calculated using the following formula [3, 7, 19], while assuming 1:1 allocation ratio:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 (p_C(1-p_C) + p_T(1-p_T))}{(p_C - p_T - (1-\lambda)M_1)^2}, \tag{2.3}$$

where $z_{1-\alpha}$, $z_{1-\beta}$ are $1-\alpha$, $1-\beta$ quantiles of standard normal distribution respectively. Specifically, α, $1-\beta$ represents desired levels of target type-I error and power respectively. Under H_0 in 2.1 and assuming equality for the true proportions for both treatment groups $p_C = p_T$, given the same type-I error and power, the difference between sample size calculations for some value of $\lambda = \lambda^*$ and μ_λ will be proportional to $\frac{1}{(1-\lambda^*)^2 M_1^2} - \frac{1}{(1-\mu_\lambda)^2 M_1^2}$. This means that for example, if $p_C = p_T = 0.8$, $\alpha = 2.5\%$, $1-\beta = 85\%$ and $\mu_\lambda = 0.7$, the sample size per arm using (2.3) for $\lambda = \mu_\lambda$ is 593, while for $\lambda = 0.71$ it would be 634, which correspond to additional 82 subjects to be recruited to a study.

The scenario presented here, where the F and its parameters are known is of course hypothetical and cannot happen in practice. We use it in order to motivate the readers to think about the fraction of the standard treatment effect as of random variable. Next we discuss how F and it's parameters could be estimated from a survey of clinical experts.

2.2 Estimating Fraction Preservation Though a Survey

The distribution F and it's parameters μ_λ, σ^2_λ are considered unknown and ought to be estimated ideally from a clinical experts survey conducted at the design stage of the trial. We assume that in total K values of λ were collected from clinicians: $\lambda_1, \dots, \lambda_K$.

Assuming independence between the clinical expert survey data and the outcome variable in the non-inferiority trial, a maximum likelihood estimates of p_C, p_T, and μ_λ

are $\hat{p}_C = \frac{1}{N_C}\sum_{j=1}^{N_C} Y_{Cj}$, $\hat{p}_T = \frac{1}{N_T}\sum_{j=1}^{N_T} Y_{Tj}$ and $\hat{\mu}_\lambda = \frac{1}{K}\sum_{k=1}^{K}\lambda_k$ respectively.

Given a sufficiently large sample size per treatment arm, the following approximate result holds:

$$\hat{p}_C - \hat{p}_T \sim N\left(p_C - p_T, \frac{p_C(1-p_C)}{n_C} + \frac{p_T(1-p_T)}{n_T}\right), \quad (2.4)$$

where $\frac{p_C(1-p_C)}{n_C} + \frac{p_T(1-p_T)}{n_T}$ is the variance term, that can be estimated by replacing p_C, p_T with $\hat{p}_C$, $\hat{p}_T$ respectively.

Similarly, for a sufficiently large clinical experts survey, the following approximate result holds too:

$$\hat{\mu}_\lambda \sim N\left(\mu_\lambda, \frac{\sigma_\lambda^2}{K}\right), \quad (2.5)$$

where the variance term can be estimated by $\hat{\sigma}_\lambda^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\lambda_k - \hat{\mu}_\lambda)^2$.

Using the above derivations, one can test the hypothesis in (2.2) at α level, by comparing the bound UB of the upper $(1-\alpha)100\%$ CI with zero:

$$UB = \hat{p}_C - \hat{p}_T - (1-\hat{\mu}_\lambda)M_1 + z_{1-\alpha}\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_C} + \frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{M_1^2\hat{\sigma}_\lambda^2}{K}}. \quad (2.6)$$

If the quantity in (2.6) is smaller than zero, the null hypothesis in (2.2) will be rejected and the new treatment will be declared non-inferior to the standard of care. This approach is in essence synthesis of the information between clinical experts opinions and the data in a new non-inferiority trial. It corresponds to an objective determination of new treatment's non-inferiority, as it takes into account opinions of the multiple clinical experts and the variability associated with such.

The apparent issue with the above approach is that in practice, it is reasonable to assume that K is small. Therefore the sample of the observed clinical experts responses might not be representative of the clinical experts population, and the normal approximation in (2.5) may not hold.

Although it might be challenging to survey a large number of clinicians to obtain their opinion about λ, other information related to clinical experts opinions could be more accessible (for example, number of years of treating a disease of interest or number of patients treated), and will be determined as X for the rest of this paper. In general, X can be a vector, here for simplicity we will assume that it contains only one random variable. As a result we have a dataset which contains a fully observed X and a partially observed λ. This resembles a missing data problem, which is thoroughly discussed in the next section.

2.3 Treating Fraction Preservation as Missing Data

Observing all the values of λ from a representative sample of experts would be extremely helpful and would allow proper use of (2.5), however such observation is unlikely to happen in practice. As a result we propose to treat unobserved values of λ as missing information. Given additional variable X, which is observed for all the experts from a representative sample, we can use MI procedure to properly estimate μ_λ and σ_λ, which can then be used in (2.6).

For MI purposes, we define a quantity of interest $Q_\lambda = \mu_\lambda$. We assume that for completely observed values of λ, $(Q_\lambda - \hat{Q}_\lambda) \sim N(0, U_\lambda)$, where $\hat{Q}_\lambda$ is an estimate of Q_λ and U_λ is the variance of $(Q_\lambda - \hat{Q}_\lambda)$. Following a maximum likelihood approach: $\hat{Q}_\lambda = \hat{\mu}_\lambda$ and $U_\lambda = \frac{\hat{\sigma}_\lambda^2}{K}$.

Following the classification and regression trees (CART) imputation method developed by [4], we use completely observed values of X to impute the incomplete data L times. CART was chosen over a normal model imputation model [30] due to its tendency to produce small mean squared errors [1]. The imputations were produced using multiple imputation chained equations (MICE) [37]. As a result we have L completed datasets, from which we calculate L pairs of estimates $(\hat{Q}_\lambda^{(l)}, U_\lambda^{(l)})$, $(l = 1, \ldots, L)$. Using Rubin's rules, we can then combine the L pairs of estimates to receive the overall point estimate $\bar{Q}_\lambda = \frac{1}{L}\sum_{l=1}^{L}\hat{Q}_\lambda^{(l)}$, and variance estimate $T_\lambda = \bar{U}_\lambda + (1+\frac{1}{L})B_\lambda$, where $\bar{U}_\lambda = \frac{1}{L}\sum_{l=1}^{L}U_\lambda^{(l)}$ is within imputation variance, and $B_\lambda = \frac{1}{L-1}\sum_{l=1}^{L}(\hat{Q}_\lambda^{(l)} - \bar{Q}_\lambda)^2$ is between imputation variance. Following this procedure, we have $(Q_\lambda - \bar{Q}_\lambda)/\sqrt{T_\lambda} \sim t_{\nu_\lambda}$, where $\nu_\lambda = (L-1)(1 + \frac{\bar{U}_\lambda}{B_\lambda(1+1/L)})^2$.

If the subject level data is fully observed, the $\hat{\mu}_\lambda$ and $\frac{\hat{\sigma}_\lambda^2}{K}$ in (2.6) are then replaced with $\bar{Q}_\lambda$ and T_λ respectively. In addition the $z_{1-\alpha}$ in (2.6) is replaced with an appropriate cut-off value from a sum of normal and Student's t-distribution using general purpose convolution algorithm with Fast Fourier Transformation (FFT) [20, 31].

When the subject level data are incomplete, a separate MI procedure should be applied for that data. For simplicity we assume that the incomplete data follow ignorable missingness. Ignorable missingness is based on the following two assumptions: 1) the incompleteness of the subject level data was either completely random or related to the observed study information, and 2) the parameter of interest in the NI trial is independent from the missingness process parameter [21, 35]. We will now define an additional quantity of interest $Q_Y = p_C - p_T$, so that for completely observed data $(Q_Y - \hat{Q}_Y) \sim N(0, U_Y)$, where $\hat{Q}_Y = \hat{p}_C - \hat{p}_T$ and $U_Y = U_C - U_T$ with $U_i = \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$ for $i = C, T$. Using a logistic regression model with MICE and observed covariates, the incomplete data is imputed D times. Similarly to the margin imputation described above, we will end up with D pairs of estimates $(\hat{Q}_Y^{(d)}, U_Y^{(d)})$, $(d = 1, \ldots, D)$, which then can be used in Rubin's rules to calculate $\bar{Q}_Y$ and T_Y following similar steps as described above for the margin imputation. As a result we have: $(Q_Y - \bar{Q}_Y)/\sqrt{T_Y} \sim t_{\nu_Y}$, where ν_Y has a similar form as ν_λ above. Now, in addition to replacing $\hat{\mu}_\lambda$ and $\frac{\hat{\sigma}_\lambda^2}{K}$ with $\bar{Q}_\lambda$ and T_λ in (2.6) respectively, we will

also replace the $\hat{p}_C - \hat{p}_T$ and $\frac{\hat{p}_C(1-\hat{p}_C)}{n_C} + \frac{\hat{p}_T(1-\hat{p}_T)}{n_T}$ in (2.6) with $\bar{Q}_Y$ and T_Y respectively. Also the $z_{1-\alpha}$ is replaced with an appropriate cut-off value from a sum of two Student's t distribution using the FFT algorithm.

2.4 Rates of Missing Information

Schafer [32] recommends calculating rates of missing information, while pointing out that such quantities could be useful when evaluating the effect of the incomplete data on the inferential uncertainty of the parameter of interest. In our case the missingness is due to unobserved clinical experts opinions regarding λ, as well as due to unobserved subject level data when the patient data are incomplete.

We estimated rates of missing information due to unobserved λ as: $\gamma_\lambda = \frac{B_\lambda}{B_\lambda + U_\lambda}$, and rates of missing information due to unobserved subject level data as $\gamma_Y = \frac{B_Y}{B_Y + U_Y}$ [11]. Since, we assume that the two data sources are independent, and the MI is done for each dataset separately, rather than conditionally, the total rate of missing information was defined as $\gamma = \gamma_\lambda + \gamma_Y$.

2.5 Simulations Details

2.5.1 Subject Level Information Is Fully Observed

Suppose the overall population of physicians consists of 1000 medical doctors (MDs), who treat a specific condition. Further, we assume that 300 of these MDs, representative of the overall population, come to a clinical conference ($K = 300$), and that it is feasible for us to survey only 3% of them (9 MDs). Also, we assume that years of experience treating the condition are known for all the MDs, who come to the conference.

Following the above notation, λ_k is a fraction preservation of the control treatment effect over placebo for kth clinical expert, also let X_k be a number of years that clinical expert has been treating a condition of interest. Without loss of generality we will drop the index k from the following explanation. Assume that for any $(\lambda, X) \sim N_2(\mu_\lambda = 0.7, \mu_X = 20, \sigma_\lambda = 0.12, \sigma_X = 7, \rho)$, where $\rho \in (0.4, 0.7)$. The positive correlation between X and λ indicates that more experienced clinical experts are prone to be more conservative with respect to the clinical margin choice. For brevity and due to similarity between the results, we only present results for $\rho = 0.4$. Let R_{λ_k} be an indicator variable for whether λ_k was observed ($R_{\lambda_k} = 1$ means that clinician k did not participate in the survey). Two scenarios of participation were considered: more experienced clinicians are more likely to participate in the survey, and a random sample from the K clinicians above. For the first scenario, the observed/unobserved values of λ were assigned using $P(R_{\lambda_k} = 1 | X > 20) = 0.95$ and $P(R_{\lambda_k} = 1 | X \leq 20) = 0.99$, while for the second scenario $P(R_{\lambda_k} = 1 | X > 20) = P(R_{\lambda_k} = 1 | X \leq 20) = 0.97$.

The value of M_1 was set to be 0.23 which was assumed to be known from a meta-analysis of the relevant historical trials. In addition the subject level data was generated using a combination of $p_C = 0.8$, $p_T \in (0.775, 0.8, 0.825)$ and $n_C = n_T \in (250, 500)$, which resulted in total of 6 scenarios. The values considered for the simulation are partially based on completed NI trials [8, 9]. Each scenario was simulated 5000 times, i.e. both MDs population sample and NI trial data were simulated 5000 times.

As stated previously non-inferiority of the new treatment was determined using confidence interval in (2.6). The NI decision was considered objective (OBJ) if it was based on the representative sample of MDs (300 MDs). Other methods used for NI decision were: MI of the margin as described in the previous section with X and $L = 20$, using only observed λ values from the survey (OBS) (only 9 MDs), as well as minimum and maximum values of λ from the representative sample of the K clinicians (MIN and MAX respectively) (one MD each). Minimum and maximum values were considered in order to demonstrate how the NI decision could be affected by consulting only one MD during the conference, who happens to be the least or the most conservative clinician in that conference.

The methods' performances were assessed by comparing the rates of the NI decision to the OBJ decision rate. A decision rate was calculated as a proportions of times NI was inferred out of the 5000 simulations. The most favorable approach is the approach, for which the NI decision rate is the closet to the OBJ NI decision.

2.5.2 Subject Level Information Is Incomplete

After comparing NI decision rates as described in the previous section, where the subject level information was considered completely observed, we turn to evaluation of NI decision rates when such information is incomplete. For the purposes of this evaluation, we only used survey data where the more experienced MDs were more likely to participate in the survey, a situation that is likely to appear in practice. The incomplete primary outcome data was assumed to follow ignorable missingness, including missing completely at random (MCAR) and missing at random (MAR) processes [29].

In order to impose both MCAR and MAR processes, a variable Z was added to the NI trial simulation. Z was set to have higher values for control treatment group and have higher values for subjects experiencing an event of interest in both groups. Specifically, $Z|C, Y = 1 \sim N(180, 20)$, $Z|C, Y = 0 \sim N(100, 20)$, $Z|T, Y = 1 \sim N(130, 20)$, $Z|T, Y = 0 \sim N(80, 20)$. Z could be seen as a patient reported outcome (PRO) measured during the study, and is positively correlated with the outcome of interest.

Let R_{Sij} be an indicator variable for whether Y_{ij} was observed ($R_{Sij} = 1$ means that outcome Y_ij was unobserved for patient j in treatment i). The following logistic function was used to determine observed/unobserved values of Y in each treatment group:

$$P(R_{Sij} = 1) = \frac{1}{1 + \exp(-\theta_0 - \theta_{1i} Z_{ij})}, \qquad (2.7)$$

where $\theta_0 = \log(\frac{DO}{1-DO}) - \theta_{1i}\bar{Z}_i$, $\bar{Z}_i = \sum_{j=1}^{n_i} Z_{ij}$, θ_{1i} represents the effect of Z in group i on the missingness, and DO stands for the overall drop-out rate, which was assumed to be the same in both treatment groups and was set to 20% as a reasonable upper bound for NI trials that encounter some level of missingness [25]. The following two sets of values were considered for θ_{1i}: $\theta_{1C} = \theta_{1T} = 0$, which means that PRO measure Z didn't affect the drop-out of patient j in treatment group i, and $\theta_{1C} = -0.009$, $\theta_{1T} = 0.013$, which means that patients with lower values in Z were more likely to drop out in the control group, whereas the opposite effect was set in the new treatment group. As a result, the first set of the values for θ_{1i} specified above constituted to MCAR process, while the later represented MAR process. Following that, the difference between the two proportion $p_C - p_T$, was unbiased when estimated from the complete cases under MCAR, and biased under MAR with observed difference being more profound than it actually is.

The incomplete subject level data was multiply imputed $D = 20$ times as described in Section 2.3, and consequently used for NI decision based on MI approach. For OBS/MIN/MAX approaches the complete cases from the NI trial were used. The performance of the methods was carried out using the same evaluation criteria as presented in Section 2.5.1.

All the simulations performed here were done using R. Code is available on GitHub.[1]

3. RESULTS

For completely observed subject level data, MI approach for NI decision was shown to be the closest to the OBJ decision in most of the scenarios, with deviations between 0.14% and 4.8% (Figures 1, 2).

In general, the OBS approach was the second closet to the OBJ, with deviations of between 5.8% and 24%. This was followed by the MIN, which resulted with deviations between 3.4% and 65%. The MAX resulted in the highest deviations, that ranged between 22% and 71%.

When the subject level data was partially observed under MCAR assumption, the MI based decision was the closest to the OBJ decision in most of the scenarios, and deviated by 2% to 7.2% from the OBJ rates (Figure 3). In case where $p_T = 0.825$, $n = 500$, the MIN approach performed similar to MI. The decision rate for MIN was 100% (Table 1), which means that all of the 5000 simulated studies concluded NI of the new treatment. This result is not surprising, since in this case the new treatment is actually superior by 2.5% to a standard treatment, which means that it would be easier to claim NI. Moreover, the MIN approach represents the least conservative view of the margin, which again would make NI claim easier to make. For the rest of the scenarios, MIN had over 20% deviation from OBJ decision rates. OBS decision rates deviated between 11% and 31%, while MAX deviation ranged between 22% and 72%.

[1]See repositories: https://github.com/yuliasidi/ch2sim, https://github.com/yuliasidi/m2imp, https://github.com/yuliasidi/bin2mi

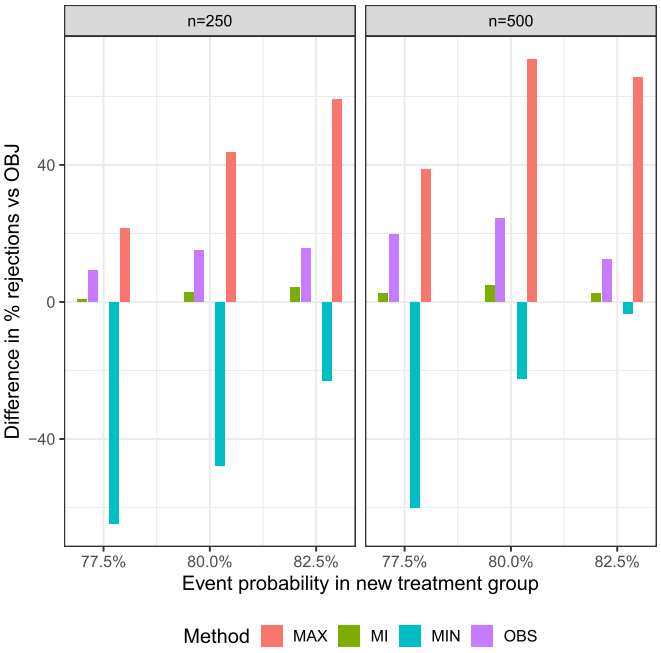

Figure 1: Deviation from objective NI decision, when more experienced MDs are more likely to participate in the survey, subject level data are fully observed.

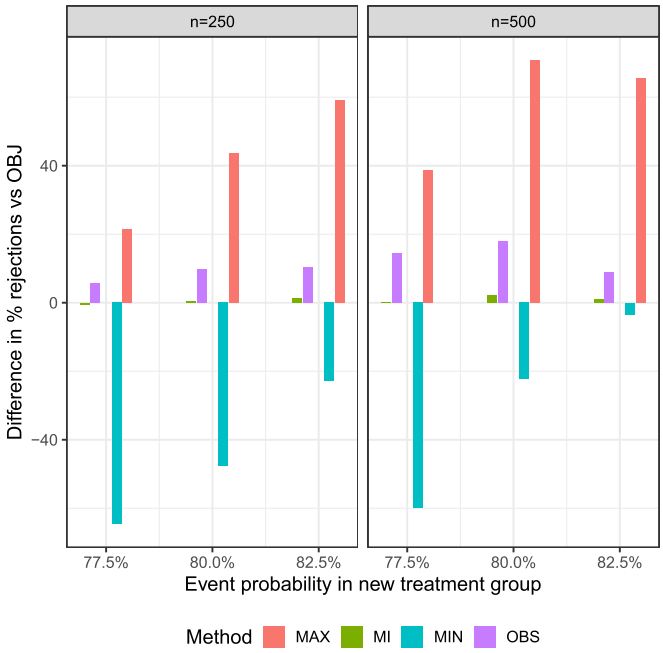

Figure 2: Deviation from objective NI decision, when MDs participation in the survey is completely random, subject level data are fully observed.

For MAR assumption for subject level data, the MI decision approach performed overwhelmingly better than the OBS and the MAX approaches (Figure 4 and Table 2).

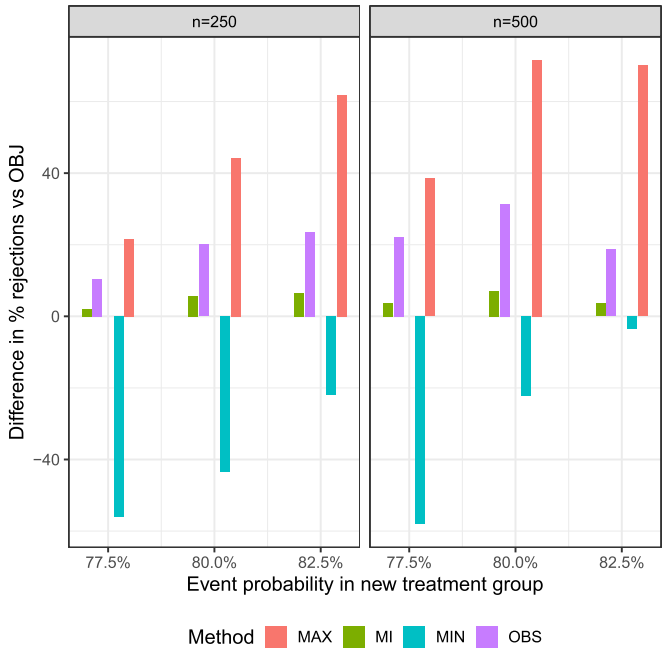

Figure 3: Deviation from population based non-inferiority decision, subject level data are MCAR.

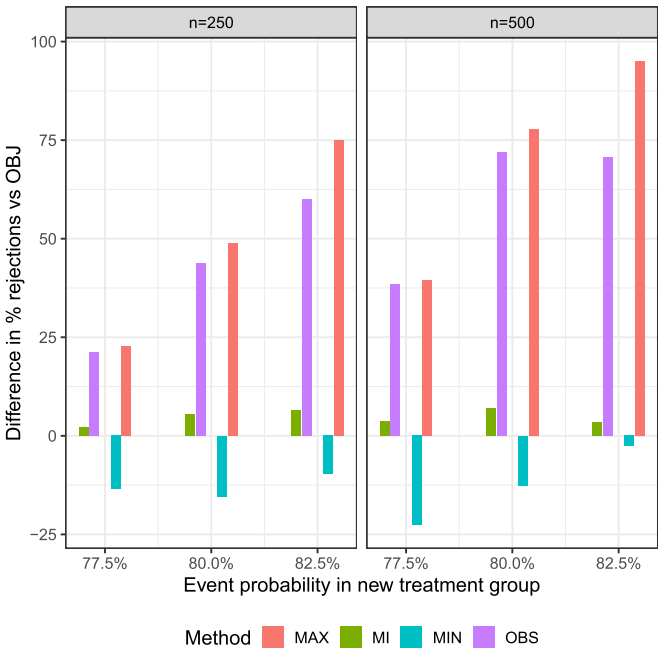

Figure 4: Deviation from population based non-inferiority decision, subject level data are MAR.

Table 1. Percent of studies concluding NI by method, when more experienced MDs are more likely to participate in the survey, subject level data are MCAR.

p_T	n	OBJ	MI	OBS	MIN	MAX
0.775	250	22.6	20.6	12.1	78.5	1.1
0.775	500	39.4	35.7	17.2	97.5	0.9
0.800	250	49.1	43.5	29.0	92.6	4.8
0.800	500	77.7	70.5	46.4	99.8	6.1
0.825	250	76.9	70.3	53.4	98.7	15.1
0.825	500	96.6	93.0	77.9	100.0	26.5

Table 2. Percent of studies with non-inferiority decision by method, subject level data are MAR.

p_T	n	OBJ	MI	OBS	MIN	MAX
0.775	250	22.6	17.5	1.4	36.1	0.0
0.775	500	39.4	29.8	1.0	62.0	0.0
0.800	250	49.1	40.1	5.3	64.4	0.2
0.800	500	77.7	66.5	5.7	90.3	0.0
0.825	250	76.9	66.4	17.0	86.6	2.0
0.825	500	96.6	91.4	26.0	99.2	1.5

Moreover, the deviations from the OBJ decision rates increased dramatically for OBS and MAX. This is reasonable, since the apparent difference in proportions for MAR is larger than it really is, which means that it is harder to claim NI. The MIN approach, however showed similar results to MI for $p_T = 0.825$ scenarios, as well as $p_T = 0.8$, $n = 250$ (Figure 4).

The rates of missing information due to unobserved λ were between 30% and 35% for both $\rho \in (0.4, 0.7)$ when more experienced MDs were more likely to participate in a survey, and between 27% and 33% when the survey participation was completely random. It should be noted that, as expected, in both cases higher rates of missing information were observed for $\rho = 0.4$. For the incomplete subject level data, the rates of missing information due to unobserved patient data ranged between 5% and 6% for both MCAR and MAR. As a result, the total rates of missing information were between 35% and 40%. As can be seen, the main contributor to the overall rates of missing information is unobserved clinical experts opinions.

4. DISCUSSION

With NI trial design being more frequently used in recent years, it is imperative to address concerns raised by several systematic reviews [39, 34, 28, 2, 25, 36]. One of the major issues that was raised in these reviews is a lack of justification for the clinically acceptable margin. A choice of the margin is critical as it directly affects the design stage of a NI study, as well as interpretation of the results once the study is complete. Even if, other common issues related to the NI design, such as availability of the historical data and the consistency of standard treatment effect over placebo are resolved, it is still not clear how to choose a clinically acceptable margin. Two reviews [28, 23] suggested using surveys to help set the non-inferiority margin, albeit using two different populations: clinical experts and patients respectively.

The selected margin is a function of the context of the trial setting (disease, current standard of care, treatment costs, side effects, etc.), and the margin selection procedure should take this context into account. Conducting a survey at a conference or symposium focused on developing and disseminating treatments for the disease under study would be an ideal setting. At a symposium, the clinical experts would be actively discussing the current standard of care, and would be actively considering the context in which a new treatment could be judged as non-inferior. Indeed, this is borne out by [26], wherein the non-inferiority margin was set more conservatively than initially due to clinical experts surveyed responses.

We presented a novel framework, where we propose to treat the margin as missing information and estimate it from a small survey of clinical experts. This framework allows an objective estimation of clinical margin and provides justification for its choice. Furthermore, within the framework we evaluated the performance of several methods by comparing the NI decision rates from each method with the objective decision rates. Overall, we found that MI was the most favorable method. Although, the least conservative margin approach had similar results to MI in several scenarios, in general, it had high deviations from the objective decision rates in other scenarios. Also, the most conservative choice of clinically acceptable margin was the least favorable method, with largest deviations from the objective decision rates. Both the most and the least conservative margin choices show the implication and risk of consulting with only one clinical expert, who might have extreme views regarding margin choice.

The rates of missing information due to the unobserved clinical experts opinions were the main contributor to the overall rates of missing information. This underlines the importance of considering uncertainty associated with the margin choice when it is observed for a small fraction of clinical experts. In addition, it has implication on a study design stage, when the allocation of study funds is discussed. Given a limited study budget, an entity running the study might consider allocating a considerable amount of funds toward the design stage, including margin determination through a clinical experts survey.

We would also like to point out several limitations of this work. First, we only considered a limited number of scenarios. If investigators have a specific scenario in mind which differs from the ones presented here, including non-ignorable missingness, they should assess it using the framework we outlined. Multiple imputation can readily be used in non-ignorable scenarios provided the imputation model considers the missingness mechanism along with sensitivity analyses [6, 41]. Second, the framework presented here is new and have not been applied previously, therefore we cannot comment towards possible logistic issues that might arose from such data collection besides the ones specified within the framework. Third, we only consider a binary outcome while time-to-event analysis and designs have become more prevalent in non-inferiority trials [38]. Our proposed methodology should be able to be extended to time-to-event analysis through similar methods as discussed in Section 2 as multiple imputation has already been applied to time-to-event analysis [40, 41].

Given the ongoing challenges with respect to NI margin choice and justification, there is a need for a new, more evidence based, and transparent approach, which takes into considerations variability in clinical experts opinions about such choice. The margin choice has direct implication on the NI decision, which is important for both drug approval and public health policy process. We believe that the above novel framework presents a simple approach, which accounts for uncertainty associated with non-inferiority margin choice. We hope that use of this framework will allow an empirical justification of margin choice, and therefore could help resolve current practical issues related to it.

FUNDING

This project was partially supported by Award Number DMS-2015320 from the National Science Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

Accepted 2 January 2024

REFERENCES

[1] Akande, O., Li, F. and Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *The American Statistician* **71**(2) 162–170. https://doi.org/10.1080/00031305.2016.1277158. MR3668704

[2] Althunian, T. A., de Boer, A., Klungel, O. H., Insani, W. N. and Groenwold, R. H. (2017). Methods of Defining the Non-Inferiority Margin in Randomized, Double-Blind Controlled Trials: A Systematic Review. *Trials* **18**(1) 107.

[3] Blackwelder, W. C. (1982). "Proving the Null Hypothesis" in Clinical Trials. *Controlled Clinical Trials* **3**(4) 345–353.

[4] Burgette, L. F. and Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology* **172**(9) 1070–1076.

[5] CHMP (2006). Committee for Medicinal Products for Human Use (CHMP) Guideline on the Choice of the Non-Inferiority Margin. *Statistics in Medicine* **25**(10) 1628. https://doi.org/10.1002/sim.3367. MR2542359

[6] Daniels, M. J. and Hogan, J. W. (2008) *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC. MR2656068

[7] Dann, R. S. and Koch, G. G. (2008). Methods for One-Sided Testing of the Difference between Proportions and Sample Size Considerations Related to Non-Inferiority Clinical Trials. *Pharmaceutical Statistics* **7**(2) 130–141.

[8] Eriksson, B. I., Dahl, O. E., Rosencher, N., Kurth, A. A., van Dijk, C. N., Frostick, S. P., Prins, M. H., Hettiarachchi, R., Hantel, S., Schnee, J. et al. (2007). Dabigatran Etexilate versus Enoxaparin for Prevention of Venous Thromboembolism after Total Hip Replacement: A Randomised, Double-Blind, Non-Inferiority Trial. *The Lancet* **370**(9591) 949–956.

[9] Eriksson, B. I., Dahl, O. E., Huo, M. H., Kurth, A. A., Hantel, S., Hermansson, K., Schnee, J. M., Friedman, R. J., Group, R. q. N. I. S. et al. (2011). Oral Dabigatran versus Enoxaparin for Thromboprophylaxis after Primary Total Hip Arthroplasty (RE-NOVATE II). *Thrombosis and Haemostasis* **105**(04) 721–729.

[10] FDA (2016). Non-Inferiority Clinical Trials.

[11] Harel, O. (2007). Inferences on Missing Information under Multiple Imputation and Two-Stage Multiple Imputation. *Statistical Methodology* **4**(1) 75–89. https://doi.org/10.1016/j.stamet.2006.03.002. MR2339010

[12] Harel, O. and Zhou, X.-H. (2007). Multiple Imputation: Review of Theory, Implementation and Software. *Statistics in Medicine* **26**(16) 3057–3077. https://doi.org/10.1002/sim.2787. MR2380504

[13] Hung, H. J. and Wang, S.-J. (2013). Statistical Considerations for Noninferiority Trial Designs without Placebo. *Statistics in Biopharmaceutical Research* **5**(3) 239–247.

[14] Hung, H. J., Wang, S.-J. and O'Neill, R. (2005). A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-Inferiority Trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **47**(1) 28–36. https://doi.org/10.1002/bimj.200410084. MR2135887

[15] Hung, H. J., Wang, S.-J. and O'Neill, R. (2007). Issues with Statistical Risks for Testing Methods in Noninferiority Trial without a Placebo Arm. *Journal of Biopharmaceutical Statistics* **17**(2) 201–213. https://doi.org/10.1080/10543400601177343. MR2345704

[16] Hung, H. J., Wang, S.-J. and O'Neill, R. (2009). Challenges and Regulatory Experiences with Non-Inferiority Trial Design without Placebo Arm. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51**(2) 324–334. https://doi.org/10.1002/bimj.200800219. MR2668686

[17] Hung, H. J., Wang, S.-J., Tsong, Y., Lawrence, J. and O'Neil, R. T. (2003). Some Fundamental Issues with Non-Inferiority Testing in Active Controlled Trials. *Statistics in Medicine* **22**(2) 213–225.

[18] ICH (2000). *International Conference on Harmonisation. Choice of Control Group and Related Issues in Clinical Trials E10.*

[19] Julious, S. A. and Owen, R. J. (2011). A Comparison of Methods for Sample Size Estimation for Non-Inferiority Studies with Binary Outcomes. *Statistical Methods in Medical Research* **20**(6) 595–612. https://doi.org/10.1177/0962280210378945. MR2866347

[20] Kohl, M., Ruckdeschel, P. and Stabla, T. (2005). General Purpose Convolution Algorithm for Distributions in S4-Classes by Means of FFT. Technical Report, Citeseer.

[21] Little, R. J. and Rubin, D. B. (2014) *Statistical Analysis with Missing Data* **333**. John Wiley & Sons. https://doi.org/10.1002/9781119013563. MR1925014

[22] Liu, Q., Li, Y. and Odem- Davis, K. (2015). On Robustness of Noninferiority Clinical Trial Designs against Bias, Variability, and Nonconstancy. *Journal of Biopharmaceutical Statistics* **25**(1) 206–225. https://doi.org/10.1080/10543406.2014.923738. MR3301347

[23] Mauri, L. and D'Agostino Sr, R. B. (2017). Challenges in the Design and Interpretation of Noninferiority Trials. *New England Journal of Medicine* **377**(14) 1357–1367.

[24] Ng, T.-H. (2008). Noninferiority Hypotheses and Choice of Noninferiority Margin. *Statistics in Medicine* **27**(26) 5392–5406. https://doi.org/10.1002/sim.3367. MR2542359

[25] Rabe, B. A., Day, S., Fiero, M. H. and Bell, M. L. (2018). Missing Data Handling in Non-Inferiority and Equivalence Trials: A Systematic Review. *Pharmaceutical Statistics* **41**(4) 815–830. https://doi.org/10.1002/sim.9251. MR4386982

[26] Radford, J., Illidge, T., Counsell, N., Hancock, B., Pettengell, R., Johnson, P., Wimperis, J., Culligan, D., Popova, B., Smith, P., McMillan, A., Brownell, A., Kruger, A., Lister, A., Hoskin, P., O'Doherty, M. and Barrington, S. (2015). Results of a Trial of PET-Directed Therapy for Early-Stage Hodgkin's Lymphoma. *New England Journal of Medicine* **372**(17) 1598–1607. https://doi.org/10.1056/NEJMoa1408648.

[27] Raghunathan, T., Berglund, P. A. and Solenberger, P. W. (2018) *Multiple Imputation in Practice: With Examples Using IVEware*. Chapman and Hall/CRC.

[28] Rehal, S., Morris, T. P., Fielding, K., Carpenter, J. R. and Phillips, P. P. (2016). Non-Inferiority Trials: Are They Inferior? A Systematic Review of Reporting in Major Medical Journals. *BMJ Open* **6**(10) 012594.

[29] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* **63**(3) 581–592. https://doi.org/10.1093/biomet/63.3.581. MR0455196

[30] Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys* **81**. John Wiley & Sons. MR2117498

[31] Ruckdeschel, P., Kohl, M., Stabla, T. and Camphausen, F. (2006). S4 Classes for Distributions. *R News* **6**(2) 2–6.

[32] Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC. https://doi.org/10.1201/9781439821862. MR1692799

[33] Schafer, J. L. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research* **8**(1) 3–15.

[34] Schiller, P., Burchardi, N., Niestroj, M. and Kieser, M. (2012). Quality of Reporting of Clinical Non-Inferiority and Equivalence Randomised Trials – Update and Extension. *Trials* **13** 214. https://doi.org/10.1186/1745-6215-13-214.

[35] Sidi, Y. and Harel, O. (2018). The Treatment of Incomplete Data: Reporting, Analysis, Reproducibility, and Replicability. *Social Science & Medicine* **209** 169–173.

[36] Sidi, Y. and Harel, O. (2021). Noninferiority Clinical Trials With Binary Outcome: Statistical Methods Used in Practice. *Statistics in Biopharmaceutical Research* **13**(4) 476–482. https://doi.org/10.1080/19466315.2020.1796780.

[37] van Buuren, S. and Groothuis- Oudshoorn, K. (2010). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 1–68.

[38] VanderBeek, B. L., Ying, G.-S. and Hubbard, R. A. (2021). Survival Analysis vs Longitudinal Modeling With Multiple Imputation—A False Dichotomy. *JAMA Ophthalmology* **139**(5) 588. https://doi.org/10.1001/jamaophthalmol.2021.0508.

[39] Wangge, G., Klungel, O. H., Roes, K. C., De Boer, A., Hoes, A. W. and Knol, M. J. (2010). Room for Improvement in Conducting and Reporting Non-Inferiority Randomized Controlled Trials on Drugs: A Systematic Review. *PLoS One* **5**(10) 13550.

[40] White, I. R. and Royston, P. (2009). Imputing Missing Covariate Values for the Cox Model. *Statistics in Medicine* **28**(15) 1982–1998. https://doi.org/10.1002/sim.3618. https://doi.org/10.1002/sim.3618. MR2750806

[41] Zhao, Y., Herring, A. H., Zhou, H., Ali, M. W. and Koch, G. G. (2014). A Multiple Imputation Method for Sensitivity Analysis of Time-to-Event Data with Possibly Informative Censoring. *Journal of Biopharmaceutical Statistics* **24**(2) 229–253. https://doi.org/10.1080/10543406.2013.860769. MR3196139

Yulia Sidi. Merck & Co.

Benjamin Stockton. Department of Statistics, University of Connecticut. E-mail address: benjamin.stockton@uconn.edu

Ofer Harel. Department of Statistics, University of Connecticut. E-mail address: ofer.harel@uconn.edu