Contemporary Clinical Trials Communications 18 (2020) 100567

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: http://www.elsevier.com/locate/conctc

Incomplete data analysis of non-inferiority clinical trials: Difference between binomial proportions case

Yulia Sidi, Ofer Harel *

Department of Statistics, University of Connecticut, USA

ARTICLE INFO

Keywords:
Binary outcome
Incomplete data analysis
Multiple imputation
Non-inferiority design

ABSTRACT

Background: Incomplete data analysis continues to be a major issue for non-inferiority clinical trials. Due to the steadily increasing use of non-inferiority study design, we believe this topic deserves an immediate attention.
Methods: We evaluated the performance of various strategies, including complete case analysis and various imputations techniques for handling incomplete non-inferiority clinical trials when outcome of interest is difference between binomial proportions. Non-inferiority of a new treatment was determined using a fixed margin approach with 95-95% confidence interval method. The methods used to construct the confidence intervals were compared as well and included: Wald, Farrington-Manning and Newcombe methods.
Results: We found that worst-case and best-case scenario imputation methods should not be used for analysis of incomplete data in non-inferiority trial design, since such methods seriously inflate type-I error rates and produce biased estimates. In addition, we report conditions under which complete case analysis is an acceptable strategy for missing at random missingness mechanism. Importantly, we show how two-stage multiple imputation could be successfully applied for incomplete data that follow missing not at random patterns, and thus result in controlled type-I error rates and unbiased estimates.
Conclusion: This thorough simulation study provides a road map for the analysis of incomplete data in non-inferiority clinical trials for different types of missingness. We believe that the results reported in this paper could serve practitioners who encounter missing data problems in their non-inferiority clinical trials.

1. Introduction

Non-inferiority (NI) clinical trials seek to show that efficacy of a new treatment is not considerably worse than that of a standard treatment [1]. Such minimally clinically acceptable deviation is called margin. While a portion of a standard treatment effect may be lost by a non-inferior agent, it offers other benefits, such as less severe adverse events, improved drug adherence and/or lower costs [2]. An NI trial design is considered when the use of placebo is unethical, as delaying treatment with a standard care would cause irreversible health damage or death [1,3].

As most clinical trials, NI trials are prone to have incomplete data, which if not properly analyzed might lead to bias in study results [4]. The importance of avoiding missing data, and performing appropriate analysis of incomplete data in clinical trials has been extensively discussed [5–9]. However, the missing data topic received a little attention with respect to NI trials [10–12]. Only a few simulation studies were conducted to assess impact of different analysis strategies for NI trials [12–14]. Moreover, the lack of deliberation around the missing data problem is evident in the published NI trials. Rehal et al. [15] reported that over 50% of reviewed NI trials didn't mention any imputation methods used in the statistical analysis. Similarly Rabe et al. [16], showed that 50% of the reviewed NI and equivalence articles used complete case analysis (CCA), a method that is generally known to produce biased results [4].

One of the principled approaches that can be used for a proper analysis of incomplete data is multiple imputation (MI) [17]. In this paper we evaluate the performance of two-stage MI, an extension of a conventional MI method [18–20] along with CCA and best/worst-case imputation methods for analysis of incomplete NI data. Specifically, we focus on NI trials assessing difference between binomial proportions, a commonly used outcome of interest [16]. In-line with FDA's recommendation to use confidence intervals (CIs) to test NI [1], we consider the following commonly used methods for a straightforward construction of CI for a difference between binomial proportions: Wilson-Newcombe (WN) [26] Farrington and Manning (FM) [25], and

* Corresponding author.
E-mail address: ofer.harel@uconn.edu (O. Harel).

https://doi.org/10.1016/j.conctc.2020.100567
Received 25 November 2019; Received in revised form 3 April 2020; Accepted 11 April 2020
Available online 4 May 2020

Wald [21]. Following a thorough simulation study while implementing different missingness mechanisms [4,22], we provide recommendations regarding the above incomplete data analysis strategies.

According to the recent "Estimands and Sensitivity Analysis in Clinical Trials" (ICH E9(R1)) guideline, handling of the intercurrent events, such as treatment discontinuation is embedded in the estimand's description [23]. Specifically, the guideline states that occurrence of the intercurrent events in the NI trials using treatment policy strategy might falsely contribute to apparent similarities between the treatment groups, and therefore requires a "careful reflection" [23]. We believe that, our work provides a useful road map regarding a proper handling of incomplete data for NI trials, and thus could be helpful in addressing the above regulatory warning.

In Section 2, we introduce CI methods mentioned above, the general missing data framework, and two-stage MI strategy. In Section 3, we present the simulation set-up, followed by the results in Section 4, and conclusions in Section 5.

2. Methods

2.1. Confidence intervals for difference between proportions

Let's assume that the primary endpoint in our trial is a difference between proportions of favorable events in control (C) and new treatment (T) groups. Let $Y_{ij} \sim Bernoullli(p_i)$ be an event indicator for subject j, $(j = 1, \ldots, n_i)$ in treatment group i $(i = C, T)$, where n_i is the total number of subjects in group i and $Y_{ij} = 1$ means that the subject experienced a favorable event. If p_i is a true proportion of favorable events in group i, and the acceptable margin is Δ then the hypothesis of interest has the following form:

$$H_0 : p_C - p_T \geq \Delta \; vs \; H_1 : p_C - p_T < \Delta \, . \tag{1}$$

We assume that the fixed margin approach is used for the above hypothesis testing, i.e., the margin is specified based on the relevant historical data prior to the current NI trial [1,24]. Further, we assume that H_0 in (1) is rejected at the pre-specified α level if the upper bound of the $100(1-\alpha)\%$ CI for $p_C - p_T$ is below Δ [1].

Using a maximum likelihood approach, the proportions of favorable events in each treatment group are estimated by the average number of events in each group, and are denoted as $\widehat{p}_C$ and $\widehat{p}_T$. Let $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of a standard normal distribution, the approximate $100(1-\alpha)\%$ CI for $p_T - p_C$ using Wald method has the following form:

$$\widehat{p}_C - \widehat{p}_T \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_T(1-\widehat{p}_T)}{N_T} + \frac{\widehat{p}_C(1-\widehat{p}_C)}{N_C}} \, . \tag{2}$$

FM method has a similar form to that of Wald's CI, with only difference at the variance term estimation, where $\tilde{p}_C, \tilde{p}_T$ are maximum likelihood estimates of p_C, p_T respectively under the restriction of the null hypothesis in (1) [25]:

$$\widehat{p}_C - \widehat{p}_T \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_T(1-\tilde{p}_T)}{N_T} + \frac{\tilde{p}_C(1-\tilde{p}_C)}{N_C}} \, . \tag{3}$$

Finally, WN method is based on the Wilson's score method for a single proportion [26,27]. Let L, U be a lower and an upper $100(1-\alpha)\%$ CI bounds for $p_C - p_T$ respectively, defined as:

$$L = \widehat{p}_C - \widehat{p}_T - \sqrt{(\widehat{p}_C - l_C)^2 + (u_T - \widehat{p}_T)^2} \, , \tag{4}$$

$$U = \widehat{p}_C - \widehat{p}_T + \sqrt{(u_C - \widehat{p}_C)^2 + (\widehat{p}_T - l_T)^2} \, , \tag{5}$$

where

$$[l_C, u_C] = \left(\widehat{p}_C + \frac{z^2_{\alpha/2}}{2N_C} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_C(1-\widehat{p}_C)}{N_C} + \frac{z^2_{\alpha}}{4N^2_C}} \right) \Bigg/ \left(1 + \frac{z^2_{\alpha/2}}{N_C} \right) , \tag{6}$$

$$[l_T, u_T] = \left(\widehat{p}_T + \frac{z^2_{\alpha/2}}{2N_T} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_T(1-\widehat{p}_T)}{N_T} + \frac{z^2_{\alpha}}{4N^2_T}} \right) \Bigg/ \left(1 + \frac{z^2_{\alpha/2}}{N_T} \right) . \tag{7}$$

2.2. Missing data framework

2.2.1. Missing data assumptions

A common framework for missing data is based on the following missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [4,22]. MCAR essentially means that the missing values in the study are completely random, and independent of the data observed or not observed in the study, MAR implies that the missing values depend on observed data, and MNAR means that the missing values depend on unobserved data. Since MCAR is unlikely to hold in clinical trials [9], analysis based on this assumption should be avoided. In addition to missingness mechanism, distinctness between data model parameters and parameter involved in generation of missing values plays a central role in incomplete data analysis. For the likelihood- and Bayes-based inferences, ignorability is the weakest, most general condition which allows ignoring the missingness model. It is characterized by both MAR and distinctness between the parameters mentioned above. As a result, non-ignorability holds when at least one of these two assumptions is violated. A detailed review of missingness mechanisms, ignorability, and the relation between these could be found elsewhere [19,28,29]. For simplicity, in this paper we will use MAR/ignorabile and MNAR/non-ignorabile terms interchangeably.

2.2.2. Missing data methodology

MI could be applied for any type of missingness structure, including MNAR. A detailed review and implementation of MI could be found elsewhere [28,30]. When data are MNAR, a missingness model needs to be specified. In practice, an exact specification of such model is difficult, if not impossible, as it relies on a set of unverifiable assumptions. Thus, the imputation model could be considered missing, and be multiply imputed together with subject level data using two-stage MI [19,20]. This approach allows to incorporate uncertainty associated with both, choice of the imputation model, and imputed subject level data into the final inference using simple arithmetic combination rules [18–20,31].

It is well known that while CCA generates unbiased estimates under MCAR, it is generally not the case for MAR [4]. Conventional MI on other hand produces unbiased results under both MCAR and MAR [4], and therefore is usually recommended over CCA. Despite this, there are still certain conditions under which CCA would result in unbiased estimates under MAR and therefore could be safely used [32]. The advantage of conventional MI over CCA for NI trials assessing difference between binomial proportions under MAR was previously shown in terms of unbiasedness and control of the type-I error [13]. The authors, however, did not evaluate cases in which CCA provides unbiased estimates for the treatment effect. Therefore, we explore such conditions here. In addition, conventional MI may result in biased estimates under MNAR, unless, relevant auxiliary variables are included in the imputation model [33,34]. The inflation of type-I error for NI trials under MNAR, when analyzed with conventional MI was reported in [13]. To resolve the issue of type-I error inflation, and consequently treatment effect bias for NI trials under MNAR, we proposed to use the two-stage MI procedure described in the following section.

2.3. Two-stage multiple imputation

If, for example, we knew that the event probability of missing values is 10% greater than the event probability in the observed values, then we

could easily specify an imputation model to account for that. The imputation model could be based on a simple transformation of ignorable imputed values to non-ignorable ones using a multiplier k [17]. More specifically we can adjust ignorable imputed probability of event ($\hat{p}^{ign}$), to a non-ingorable imputed probability of event ($\hat{p}^{nonign}$) as follows: $\hat{p}^{nonign} = k^{*}\hat{p}^{ign}$. Unfortunately, in practice it is almost impossible to make a statement similar to the above 10% example, and consequently set up one particular value of k with an absolute confidence. Therefore, following previous work by Siddique et al. [19,20] we suggest to specify a distribution for k, which corresponds to the imputation model distribution.

In practice, the imputation model distribution needs to be specified by either a study team, or by experts who collected the data. Such distribution represents the study team's belief regarding the magnitude of the bias in the observed rate in treatment group i and their confidence in this belief. These two values represent the center of the missingness model distribution (μ_{ki}) and its variance (σ^2_{ki}), respectively. For example, if the team believes the study participants were more likely to drop-out due to lack of efficacy in the new treatment, then the team will anticipate that the observed rate in the new treatment is greater than the actual rate. As a result, μ_{kT} below 1 will be chosen, so that the ignorably imputed rate is closer to it's true value. If for the same study the team believes the observed rate in the control treatment is unbiased, then $\mu_{kC} = 1$ would represent such belief. As a result, there is a separate imputation model distribution for each treatment group: $k_C \sim N(\mu_{kC}, \sigma_{kC})$ and $k_T \sim N(\mu_{kT}, \sigma_{kT})$ for control and new treatment, respectively. We chose to use normality assumption on the k_i distributions for simplicity. Other distributions can easily replace the normal distribution.

After the imputation model distribution is specified, we can randomly draw M models from it. Within each of the imputed models, patient level data can be imputed D times, resulting in $M \times D$ complete datasets. Each of these complete datasets is then analyzed using a standard statistical method, such as methods presented in the previous sections. Results from the $M \times D$ analyses are then combined using nested imputation combination rules [18] described in the next section.

2.4. Two-stage multiple imputation combination rules

In order to introduce two-stage imputation rules, a notation close to that of Siddique et al. [18,20] is used. Let Q be a quantity of interest, that approximately follows Normal distribution for a completely observed data, i.e., $(Q - \widehat{Q}) \sim N(0, U)$, where $\widehat{Q}$ is a complete data statistic estimating Q, and U is a complete data statistic for the variance of $Q - \widehat{Q}$. The $M \times D$ imputations mentioned above correspond to $M \times D$ completed datasets, where $\left(\widehat{Q}^{(m,d)}, U^{(m,d)}\right)$ represent estimate and variance of Q, respectively from a d^{th} imputed datasets under model m ($m = 1,\ldots,M; d = 1,\ldots,D$).

Let $\overline{Q}$ be the overall mean of the $M \times D$ estimates: $\overline{Q} = \frac{1}{MD}\sum_{m=1}^{M}\sum_{d=1}^{D}\widehat{Q}^{(m,d)}$, and let $\overline{Q}_m$ be the mean of the estimated from the mth model: $\overline{Q}_m = \frac{1}{D}\sum_{d=1}^{D}\widehat{Q}^{(m,d)}$.

Also, let $\overline{U}, W, B$ be the three sources of variability, defined as the overall mean of the associated variance estimates, within-model and between model variance terms respectively. Specifically:

$$\overline{U} = \frac{1}{MD}\sum_{m=1}^{M}\sum_{d=1}^{D}U^{(m,d)},$$

$$W = \frac{1}{M(D-1)}\sum_{m=1}^{M}\sum_{d=1}^{D}\left(\widehat{Q}^{(m,d)} - \overline{Q}_m\right)^2,$$

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\overline{Q}_m - \overline{Q})^2.$$

Finally, the total variance of $Q - \widehat{Q}$ has the following form: $Tvar = \overline{U} + \left(1 + \frac{1}{M}\right)B + \left(1 - \frac{1}{D}\right)W$. The final inferences of the multiple imputed data are based on Student's t distribution $Tvar^{\frac{1}{2}}(Q - \overline{Q}) \sim t_\nu$, where ν is degrees of freedom, defined as:

$$\nu^{-1} = \left[\frac{\left(1 + \frac{1}{M}\right)B}{Tvar}\right]^2 \frac{1}{M-1} + \left[\frac{\left(1 - \frac{1}{D}\right)W}{Tvar}\right]^2 \frac{1}{M(D-1)}.$$

To implement the above procedure, we set $\widehat{Q}^{(m,d)} = \hat{p}^{(m,d)}$, where $\hat{p}^{(m,d)}$ is the estimated proportion of difference between control and new treatment from nth imputation and mth model. For Wald, and FM, the value of $U^{(m,d)}$ was set to the corresponding variance term used in the method as presented under the square root in (2) and (3). For WN, $\overline{Q}$ for each treatment group was plugged into (4–7).

3. Simulations

3.1. Simulation of fully observed data

In total 30 NI clinical trials scenarios were considered. The p_C values were set to the range between 0.6 and 0.95 by increments of 0.05. The Δ values were set to: 0.05, 0.075, 0.1, 0.15 and 0.2. All possible combinations of the above margins (Δ) and probabilities (p_C) were used, excluding cases where margin was greater or equal to the corresponding failure rate ($1 - p_C$). A margin equal to the corresponding failure rate would mean that the usage of a new treatment doubles a failure rate of the treated condition. Therefore, a margin greater or equal to the corresponding failure rate, was redefined as half of the original margin. Due to the high volume of the results, we present here only 9 of the 30 scenarios (unless stated otherwise), which are representative of the rest of the results. In addition, we assumed a one-sided type-I error of 2.5%, power of 90%, and 1:1 group allocation ratio.

Since different methods for comparison of binomial proportions might require different sample sizes [35], sample sizes were calculated for each method separately using above scenarios assumptions. For Wald and FM methods, the sample size calculations were performed by inversion of the corresponding CI formulas [35], while sample sizes for WN were estimated based on 5000 simulations. As a result the sample size per arm (n)ranged between 98 and 2017 patients.

The outcome variable Y (subscripts are omitted for simplicity) was simulated for each subject using a logistic function of treatment ($T = 0$ for control treatment, $T = 1$ for the new treatment) and two continuous baseline covariates (X_1, X_2) as follows:

$$P(Y=1) = \left[1 + e^{-\left(\alpha_y + \beta_1 * X_{1ij} + \beta_2 * X_{2ij} + \beta_{Tout} * T_{ij}\right)}\right]^{-1}. \quad (8)$$

Further details regarding parameters setting in the above model are provided in the supplemental material. The total number of simulated trials per scenario and method under each hypothesis was set to 10,000 repetitions.

3.2. Simulation of incomplete data

Let R_{ij} be a missing indicator variable for outcome Y_{ij}, such that $R_{ij} = 1$ indicates that the outcome for patient j in group i is missing while $R_{ij} = 0$ means that the outcome for that patient is observed. Upon a generation of the complete datasets, the missing outcome values were imposed using the following logistic function (subscripts are omitted for simplicity):

$$P(R=1)=\left[1+e^{-\left(\alpha+\beta_T{}^*T+\beta_Y{}^*Y+\beta_{TY}{}^*T{}^*Y+\beta_{X_2}{}^*X_2\right)}\right]^{-1}. \quad (9)$$

Parameters $\beta_T, \beta_Y, \beta_{TY}, \beta_{X_2}$ represent effects of treatment group, outcome, treatment group by outcome interaction and baseline covariate X_2 on missingness, respectively. In order to impose a specific missingness mechanism (MCAR, MAR and MNAR), different parameter values were used. The overall drop-out rates were set to 5%, 10%, 15% and 20%. The 20% drop-out rate was chosen as an upper bound, because 86% of NI and equivalence trials with incomplete data reported to have drop-out rates of up to 20% [16].

For MCAR, all model parameters but α were set to 0 ($\alpha = -log\left(\frac{1}{DO}-1\right)$, DO is a target drop-out rate). For MAR, β_{X_2} was set to $\beta_{X_2} = 1.5$, while β_T ranged between −0.9 and 0.9 in order to assess unbalanced levels of drop-out rates of 5–15% between the treatment groups. MNAR was set up to implement scenarios where dropping out of the study is associated with either lack of efficacy in the new treatment or with overwhelming efficacy in the control treatment, therefore both β_Y, β_{TY} were set to non-zero values as follows: i) $\beta_Y = -0.4,\ \beta_{TY} = 2$ for MNAR due to lack of efficacy in the new treatment; ii) $\beta_Y = -0.8,\ \beta_{TY} = -2$ for MNAR due to overwhelming efficacy in the control treatment. These two conditions were considered for MNAR, as both would lead to the observed difference between the treatments to appear smaller than it actually is, which leads to an incorrect study conclusion.

3.3. Analysis strategies for incomplete data

The following analysis strategies were used for the analysis of incomplete data: CCA, best-case scenario, worst-case scenario, two-stage MI using multiple imputation chained equations (MICE) [36].

Both best-case and worst-case scenario strategies were employed only for MCAR missingness mechanism. It was expected that these two strategies would inflate type-I errors, since they make the two treatment groups more alike, which in turn makes it easier to reject the null hypothesis specified in (1).

For MAR missingness, it was expected, that CCA strategy would lead to approximately unbiased estimates of p_C and p_T [32]. This is due to the fact that baseline covariates were balanced and had similar effect on the missingness in (9). Thus, it was expected that CCA strategy would result with type-I errors that only slightly deviate from the desired level.

For MNAR missingness, it was expected that single value imputation methods, or CCA would produce biased results with inflated type-I error rates. Although, conventional MI might produce unbiased estimates when relevant auxiliary variables are used [33,34], our simulation set-up did not address such situation and therefore we anticipated that conventional MI would not be able to provide unbiased estimates for MNAR. In order to properly analyze the incomplete data that follow such missingness process, we used two-stage MI. Two-stage MI was compared to CCA rather than to conventional MI, due to the fact that both CCA and conventional MI ought to produce biased estimates, and because CCA is an easy and dominant approach in clinical trials.

As specified in the previous section, two MNAR situations were simulated: drop-out due to lack of efficacy in the new treatment and drop-out due to overwhelming efficacy in the control treatment. For the first situation, it was expected that the observed rate in the new treatment group will be higher than it actually is, while the observed rate in the control group will be unbiased, therefore we specified $k_T \sim N(\mu_{kT}, 0.05)$ where μ_{kT} was chosen below 1 and $k_C \sim N(1,0)$. On the contrary, in the second situation it was expected that the observed rate in the control group is lower than it actually is, while the observed rates in the new treatment will be unbiased, therefore we set $k_C \sim N(\mu_{kC}, 0.05)$, where μ_{kC} was chosen above 1 and $k_T \sim N(1,0)$.

Similar to Siddique et al. [20]; D was set to 2, and M was set to 100. The multiple imputation of the subject level data, within each imputed missingness model (randomly drawn values of k_T, k_C) was performed using MICE with the two baseline covariates specified above. We also performed sensitivity analysis for k, by specifying different values for μ_{kT}, μ_{kC}, and doubling the standard deviation.

3.4. Evaluation criteria

The Wald, FM and WN performances along with analysis strategies used to handle missing data were assessed using empirical type-I error, empirical power and mean relative bias. Type-I error was estimated by the proportion of trials that reject H_0 in (1) out of the trials simulated under $H_0 : p_C = \Delta + p_T$, and was considered appropriately controlled if it fall within $[0.9\alpha, 1.1\alpha] = [0.0225, 0.0275]$ bounds [37,38]. Power was estimated by the proportion of trials that reject H_0 in (1) out of the trials simulated under $H_1 : p_C = p_T$. A relative bias was defined under $H_0 : p_C = \Delta + p_T$ as $(\hat{p}_C - \hat{p}_T - \Delta)/\Delta$ per repetition. A result was considered unbiased if the mean relative bias fall within $[-0.1, 0.1]$ bounds. The negative bias implies that the new treatment (T) is worse than it appears, thus a non-inferiority of the new treatment may be incorrectly inferred.

The simulations were run using the R-package nibinom we developed. The package with additional code to reproduce the presented results are available here: https://github.com/yuliasidi/nibinom_apply.

4. Results

4.1. Missing completely at random

For MCAR, we present results for overall drop-out rate of 20%, as these are representative for lower drop-out rates. Also, since the three methods showed very similar results, only Wald method is presented for MCAR. As can be seen in Table 1, worst-case scenario imputation strategy produced inflated type-I error rates that were more than double of the completely observed data, along with significantly biased estimates. On contrary, CCA produced unbiased estimates with type-I errors being either within the pre-specified range or very close to it.

Due to the significant inflation of type-I error for worst-case imputation method, empirical power was calculated for CCA strategy only. As expected the power is decreasing with higher drop-out rates, dropping to 81.5% (see supplemental materials). Results for best-case scenario imputation were very similar to the worst-case scenario and are, therefore, omitted.

4.2. Missing at random

Empirical type-I errors under MAR assumption, analyzed using CCA for balanced drop-out rates were well controlled in most scenarios by the three methods (Fig. 1). In addition, this strategy resulted in unbiased estimates, while the empirical power went down to 81.7% (see supplemental materials). For unbalanced drop-out rates, as expected, CCA showed slight deviations from the desired level of the type-I error. The largest empirical type-I error was equal to 0.0419 for overall drop-out rate of 20%, when the drop-out rates between the treatment groups

Table 1
Empirical type-I errors and mean relative bias for MCAR, DO = 20%, worst-case imputation scenario and CCA strategies, Wald method.

		Type-I			Bias	
p_C	Δ	Full	Worst	CCA	Worst	CCA
0.65	0.05	0.026	0.103	0.029	−0.214	−0.019
0.65	0.10	0.027	0.093	0.028	−0.210	−0.016
0.65	0.15	0.025	0.090	0.026	−0.211	−0.015
0.75	0.05	0.025	0.079	0.026	−0.201	−0.002
0.75	0.10	0.026	0.087	0.029	−0.205	−0.004
0.75	0.15	0.023	0.084	0.025	−0.209	−0.009
0.80	0.15	0.024	0.074	0.026	−0.212	−0.011
0.85	0.05	0.023	0.066	0.024	−0.194	0.008
0.85	0.10	0.028	0.067	0.026	−0.198	0.003

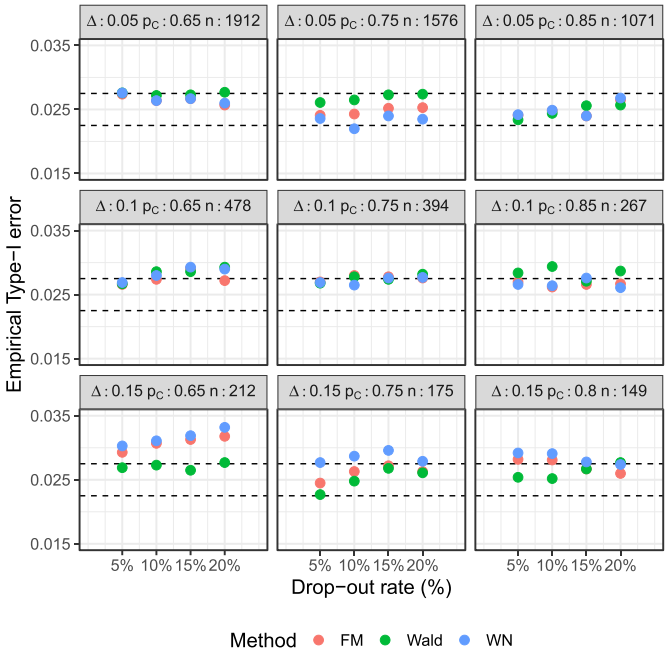

Fig. 1. Empirical type-I error CCA strategy for MAR: drop-out rates are balanced between the treatment groups.

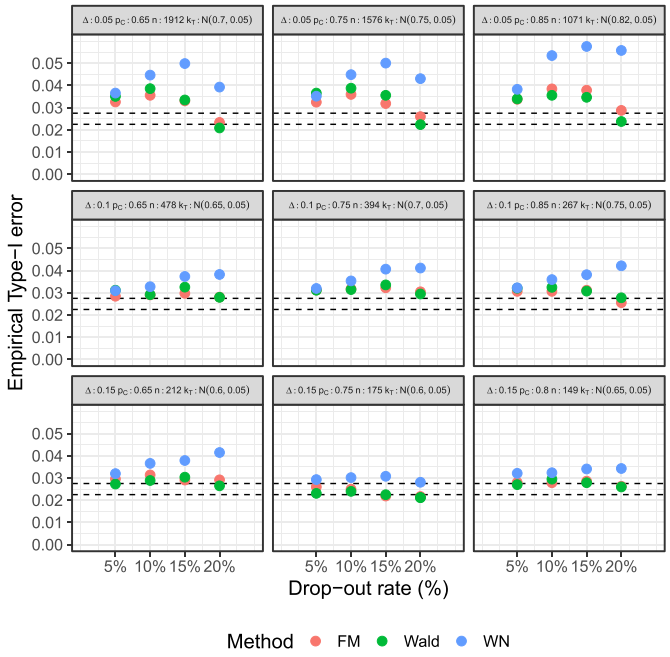

Fig. 3. Empirical type-I errors, two-stage MI strategy via MICE for MNAR due to lack of efficacy *T*.

differed by 15% (see supplemental materials). Nevertheless, the mean relative bias fall within the specified bounds for all of the scenarios, methods and drop-out rates (results are not presented).

4.3. Missing not at random

Empirical type-I errors rates for incomplete data under MNAR due to lack of efficacy in the new treatment were seriously inflated when analyzed using CCA (Fig. 2). This was not the case for two-stage MI, which produced type-I errors either within the specified bounds or very close to them (Fig. 3). In addition, for two-stage MI, WN method has shown a less favorable results when compared to Wald and FM. The advantage of two-stage MI over CCA is also demonstrated by the mean relative bias, with CCA resulting in mean relative bias as large as −0.897 for drop-out rate of 20% when using Wald method (Table 2). The corresponding mean relative bias results for the other two methods were similar to Wald and, therefore are omitted. Furthermore, while the mean relative bias was of a smaller magnitude for lower drop-out rates, CCA still resulted in biased estimates in most cases, while two-stage MI showed unbiased estimates (results not shown). The empirical power based on the two-stage MI was below the desired level of 0.9 with lowest rate of 65.8% for overall drop-out rate of 20%. This is not surprising due to variability introduced through the MI procedure (see supplemental materials). Results from MNAR due to overwhelming efficacy in the control treatment were similar in terms of type-I errors, bias and power to the MNAR due to lack of efficacy in the new treatment (see supplemental material).

In Fig. 4, we present sensitivity analysis for the choice of distribution of imputation models specified by multiplier k_T. Although type-I error rates are affected by the choice of the imputation model distribution, in all the cases the type-I errors are much smaller than the one observed for CCA strategy (solid black horizontal line).

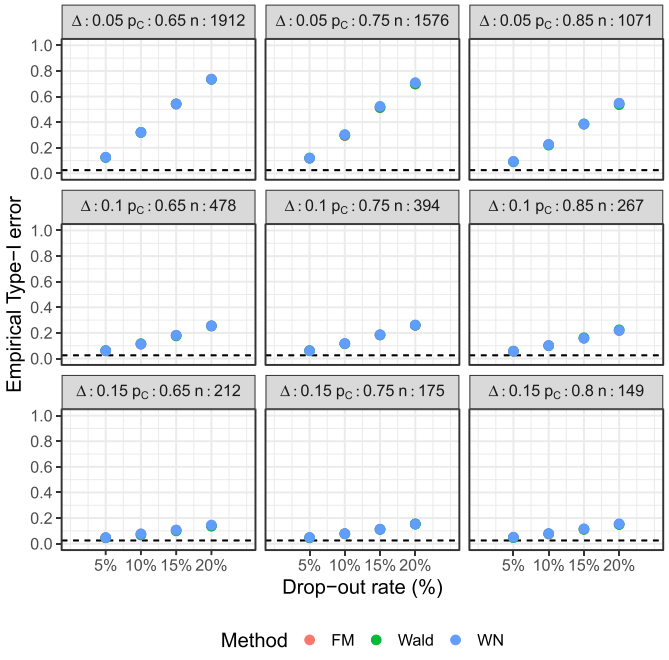

Fig. 2. Empirical type-I errors, CCA strategy for MNAR due to lack of efficacy in *T*.

Table 2
Mean relative bias for MNAR due to lack of efficacy in *T*, DO = 20%, CCA and two-stage MI strategies, Wald method.

p_C	Δ	CCA	MI
0.65	0.05	−0.897	−0.032
0.65	0.10	−0.453	0.015
0.65	0.15	−0.300	0.022
0.75	0.05	−0.852	−0.038
0.75	0.10	−0.458	0.000
0.75	0.15	−0.319	0.055
0.80	0.15	−0.328	0.019
0.85	0.05	−0.709	−0.068
0.85	0.10	−0.422	−0.001

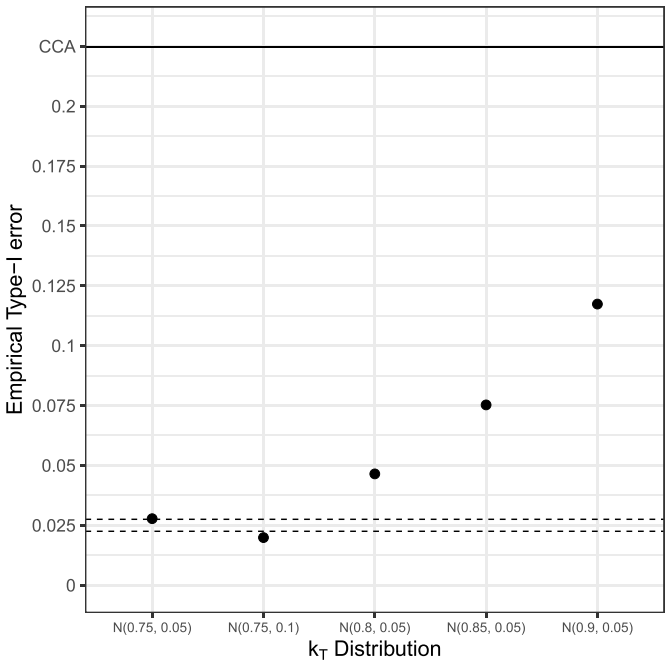

Fig. 4. Choice of different distribution parameters for k_T. Empirical type-I error, two-stage MI strategy via MICE for scenario with $p_C = 0.85, \Delta = 0.1$ for MNAR due to lack of efficacy in T.

5. Conclusion

Our work presents a thorough simulation study that assesses different strategies for analysis of incomplete data, when NI design is employed and the outcome of interest is a difference between binomial proportions. We evaluated three commonly used methods for construction of confidence intervals for the difference between binomial proportions: Wald, WN and FM.

We found that both best-case and worst-case imputation strategies perform poorly even when the incomplete data follows MCAR. This is due to the fact that by treating incomplete cases similarly for both treatment groups, we make the estimated proportions similar, which leads to erroneous conclusion of NI. According to Rabe et al. [16]; 28% of the reviewed articles that encountered some amount of incomplete data in the primary analysis, used single imputation strategy, including best/worst-case imputation. The simulation results we present here, along with the review results reported by Rabe et al. [16] are concerning. We believe that such imputation strategy should be abounded when dealing with an NI analysis.

Similar to previous work by Barlett et al. [32]; we found that CCA performs well when incomplete data follows MAR, and both baseline covariates that affect the missingness and the corresponding drop-out rates are balanced between treatment arms. In addition, when the drop-outs rates were higher in the new treatment, type-I errors might be inflated, depending on the scenario. Among cases with unbalanced drop-out rates, the highest type-I error rate that was seen is 0.0419% for overall drop-out rate of 20% with 15% higher drop-out in the new treatment. Considering the levels of inflations seen for MNAR and the fact that 0.0419% rate was reached by a relatively extreme missingness scenario, we believe that CCA could still be considered as a safe choice for MAR incomplete data. It should be noted that, if researchers assume that MAR is affected by variables that have different levels between the treatment groups, then conventional MI strategy is recommended over CCA as suggested in [13]. The importance of the findings for MAR presented here, is to demonstrate when CCA could be used and what assumptions need to be made in order to have a valid inference.

Importantly, we demonstrated that while CCA performs poorly for incomplete data under MNAR, which is also the case for conventional MI [13], two-stage MI strategy produces favorable results. We believe that, these results are of great importance for practitioners who encounter incomplete data in NI clinical trials. The limitation of this method is the specification of the distribution of the multiple imputation model, or the multiplier. Nevertheless, according to the sensitivity analysis we performed, it is clear that even if the parameters of the multiplier's distribution are shifted, the type-I error rates are still substantially lower than those seen with CCA strategy.

The results of the empirical power were in line with our expectation. In general, we saw a decrease of empirical power with increasing drop-out rates. In terms of the difference between the analysis methods considered here, we found that in most cases there was no difference between the three. However, when the two-stage MI procedure was used, WN performed less favorable than Wald and FM. This could be explained by the fact that we used a plug-in method for WN, rather than a proper MI combination rules. A method for a proper combination of multiple imputed data and analysis of difference between proportions using WN is unavailable yet and is of interest for a future research.

Although, we have looked at a variety of different scenarios, one limitation of our work is that it does not cover each possible scenario. Therefore, before finalizing statistical analysis plan for an NI trial, researchers should always consider the specific scenario they are dealing with. Another limitation of our work is that, we considered moderate to large sample sizes. We have not evaluated small sample sizes, which might require exact methods, such as the method due to Chan [39]; and thus might have different implications when applying MI strategy. In addition, only a simple transformation of ignorable to non-ignorable imputed values was evaluated for MNAR analysis. A more thorough examination of various functional forms of such transformation is of interest for future research.

In summary, we recommend employing the following analyses strategies when dealing with incomplete data for NI trials assessing difference between binomial proportions: 1) if the incomplete data follow MAR and it is reasonable to assume that the missingness is caused by balanced baseline covariates only, then CCA could be used, 2) if the data are MAR, but the missingness is caused by other unbalanced variables then following the work by Lipkovich and Weins [13] conventional MI should be used, 3) if MNAR is a more reasonable assumption, then two-stage MI should be used, 4) best-case and worst-case imputation should be avoided.

We believe that, the above recommendations are useful for practitioners who face incomplete data analysis of NI trials that assess difference between binomial proportions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2020.100567.

References

[1] FDA, Non-inferiority Clinical Trials to Establish Effectiveness; Guidance for Industry, 2016.
[2] G. Piaggio, D.R. Elbourne, S.J. Pocock, S.J. Evans, D.G. Altman, C. Group, et al., Reporting of noninferiority and equivalence randomized trials: extension of the consort 2010 statement, Jama 308 (24) (2012) 2594–2604.
[3] ICH, Choice of control group and related issues in clinical trials e10, 2000.
[4] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, vol. 333, John Wiley & Sons, 2014.
[5] L.M. Collins, J.L. Schafer, C.-M. Kam, A comparison of inclusive and restrictive strategies in modern missing data procedures, Psychol. Methods 6 (4) (2001) 330.
[6] J.D. Dziura, L.A. Post, Q. Zhao, Z. Fu, P. Peduzzi, Strategies for dealing with missing data in clinical trials: from design to analysis, Yale J. Biol. Med. 86 (3) (2013) 343.
[7] T.R. Fleming, Addressing missing data in clinical trials, Ann. Intern. Med. 154 (2) (2011) 113–117.
[8] R.J. Little, R. D'agostino, M.L. Cohen, K. Dickersin, S.S. Emerson, J.T. Farrar, C. Frangakis, J.W. Hogan, G. Molenberghs, S.A. Murphy, et al., The prevention and

treatment of missing data in clinical trials, N. Engl. J. Med. 367 (14) (2012) 1355–1360.
[9] NRC, The Prevention and Treatment of Missing Data in Clinical Trials, National Academies Press, 2011.
[10] T.R. Fleming, Current issues in non-inferiority trials, Stat. Med. 27 (3) (2008) 317–332.
[11] P. Gallo, C. Chuang-Steiny, A note on missing data in noninferiority trials, Drug Inf. J. 43 (4) (2009) 469–474.
[12] B.L. Wiens, W. Zhao, The role of intention to treat in analysis of noninferiority studies, Clin. Trials 4 (3) (2007) 286–291.
[13] I. Lipkovich, B.L. Wiens, The role of multiple imputation in non-inferiority trials for binary outcomes, Stat. Biopharm. Res. 10 (1) (2017) 57–69.
[14] B. Yoo, Impact of missing data on type 1 error rates in non-inferiority trials, Pharmaceut. Stat. 9 (2) (2010) 87–99.
[15] S. Rehal, T.P. Morris, K. Fielding, J.R. Carpenter, P.P. Phillips, Non-inferiority trials: are they inferior? a systematic review of reporting in major medical journals, BMJ open 6 (10) (2016), e012594.
[16] ICH, Estimands and Sensitivity Analysis in Clinical Trials, 2017.
[17] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, vol. 81, John Wiley & Sons, 2004.
[18] Z. Shen, Nested Multiple Imputations, Ph. D. thesis, Harvard University, 2000.
[19] J. Siddique, O. Harel, C.M. Crespi, Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial, Ann. Appl. Stat. 6 (4) (2012) 1814.
[20] J. Siddique, O. Harel, C.M. Crespi, D. Hedeker, Binary variable multiple-model multiple imputation to address missing data mechanism uncertainty: application to a smoking cessation trial, Stat. Med. 33 (17) (2014) 3013–3028.
[21] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Am. Math. Soc. 54 (3) (1943) 426–482.
[22] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[23] B.A. Rabe, S. Day, M.H. Fiero, M.L. Bell, Missing Data Handling in Non-inferiority and Equivalence Trials: A Systematic Review, Pharmaceutical Statistics, 2018.
[24] M.D. Rothmann, B.L. Wiens, I.S. Chan, Design and Analysis of Non-inferiority Trials, Chapman and Hall/CRC, 2016.
[25] C.P. Farrington, G. Manning, Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, Stat. Med. 9 (12) (1990) 1447–1454.
[26] R.G. Newcombe, Interval estimation for the difference between independent proportions: comparison of eleven methods, Stat. Med. 17 (8) (1998) 873–890.
[27] E.B. Wilson, Probable inference, the law of succession, and statistical inference, J. Am. Stat. Assoc. 22 (158) (1927) 209–212.
[28] O. Harel, X.-H. Zhou, Multiple imputation: review of theory, implementation and software, Stat. Med. 26 (16) (2007) 3057–3077.
[29] R. Yucel, Impact of the non-distinctness and non-ignorability on the inference by multiple imputation in multivariate multilevel data: a simulation assessment, J. Stat. Comput. Simulat. 87 (9) (2017) 1813–1826.
[30] J.L. Schafer, Multiple imputation: a primer, Stat. Methods Med. Res. 8 (1) (1999) 3–15.
[31] J.P. Reiter, T.E. Raghunathan, The multiple adaptations of multiple imputation, J. Am. Stat. Assoc. 102 (480) (2007) 1462–1471.
[32] J.W. Bartlett, O. Harel, J.R. Carpenter, Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression, Am. J. Epidemiol. 182 (8) (2015) 730–736.
[33] Cmph, Guideline on Missing Data in Confirmatory Clinical Trials, European Medicines Agency, London, 2010.
[34] H. Demirtas, J.L. Schafer, On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out, Stat. Med. 22 (16) (2003) 2553–2575.
[35] S.A. Julious, R.J. Owen, A comparison of methods for sample size estimation for non-inferiority studies with binary outcomes, Stat. Methods Med. Res. 20 (6) (2011) 595–612.
[36] S.v. Buuren, K. Groothuis-Oudshoorn, mice: multivariate imputation by chained equations in r, J. Stat. Software 1–68 (2010).
[37] R.S. Dann, G.G. Koch, Methods for one-sided testing of the difference between proportions and sample size considerations related to non-inferiority clinical trials, Pharmaceut. Stat. 7 (2) (2008) 130–141.
[38] P. Roebruck, A. Kühn, Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities, Stat. Med. 14 (14) (1995) 1583–1594.
[39] I.S. Chan, Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies, Stat. Med. 17 (12) (1998) 1403–1413.