The Impact of Major Events on Ongoing Noninferiority Trials, With Application to COVID-19

Brian L. Wiens and Ilya Lipkovich

ABSTRACT

The COVID-19 pandemic has impacted ongoing clinical trials. We consider particular impacts on non-inferiority clinical trials, which aim to show that an investigational treatment is not markedly worse than an existing active control with known benefit. Because interpretation of noninferiority trials requires cross-trial validation involving untestable assumptions, it is vital that they be run to very high standards. The COVID-19 pandemic has introduced an unexpected impact on clinical trials, with subjects possibly missing treatment or assessments due to unforeseen intercurrent events. The resulting data must be carefully considered to ensure proper statistical inference. Missing data can often, but not always, be considered missing completely at random (MCAR). We discuss ways to ensure validity of the analyses through study conduct and data analysis, with focus on the hypothetical strategy for constructing estimands. We assess various analytic strategies of analyzing longitudinal binary data with dropouts where outcomes may be MCAR or missing at random (MAR). Simulations show that certain multiple imputation strategies control the Type I error rate and provide additional power over analysis of observed data when data are MCAR or MAR, with weaker assumptions about the missing data mechanism.

1. Introduction

Noninferiority (NI) trials aim to demonstrate that an investigational treatment is not markedly worse than an active control treatment with known benefit. Letting the mean response rate be \( \pi_1 \) in the investigational group and \( \pi_0 \) in the control group, and assuming without loss of generality that larger values are preferred, the null hypothesis of \( H_0 : \pi_0 - \pi_1 \geq \Delta \) is tested against the alternative of \( H_a : \pi_0 - \pi_1 < \Delta \). If \( \Delta \) is chosen appropriately, rejecting the null hypothesis is tantamount to concluding not only that there is not an important difference between treatments, but also that the investigational treatment is superior to placebo. More comprehensive discussions can be found in, for example, Wellek (2010) or Rothmann, Wiens, and Chan (2012).

The recent outbreak of COVID-19 has interfered with ongoing clinical trials. Some study subjects are unable to receive treatment during the study because site personnel are unavailable to dispense or administer medications or because the sponsor is unable to ship medication. Additionally, some subjects may not be able to receive efficacy assessments due to shelter-in-place orders that leave them unable to travel to the study site. These may have heavy impact on noninferiority trials because of the need to have and to demonstrate rigorous study conduct.

Meyer et al. (2020) provided recommendations on general clinical trial data issues for studies conducted during the pandemic. In this article, we focus in particular on how sponsors can maximize the chance that a noninferiority study conducted during the COVID-19 pandemic gives useful information. In Section 2, we reinforce the importance of rigor in noninferiority studies and discuss how to maintain assay sensitivity in studies ongoing during the pandemic. In Section 3, we start with a brief discussion of handling missing data caused by pandemic in the NI context, emphasizing that methods should be determined only after choosing the estimand. Then we review methods that might be used to analyze data collected in these studies and present simulations to study the performance of several multiple imputation (MI) strategies compared with traditional nonresponder imputation and observed case (a.k.a. completer’s analyses) when the primary outcome variable is binary. We close with recommendations on imputation models to control the Type I error rate and bias while providing high power.

2. Importance of Rigor in Noninferiority Trials

In general, noninferiority trials need to be conducted with great rigor. Rigor, in this context, refers to conducting the trial according to general clinical trial principles (International Conference on Harmonisation 1998; Food and Drug Administration 2016) and also according to the written study protocol.

2.1. Assay Sensitivity

To convincingly demonstrate that two treatments are similar, it is required that a clinical trial maintain assay sensitivity (Food and Drug Administration 2016). In the context of noninferiority trials, assay sensitivity means that a trial can distinguish two treatments as different, if they have an important difference. Any study conduct that diminishes assay sensitivity, or the
appearance of assay sensitivity, will reduce confidence in a non-inferiority conclusion.

Specific aspects of the COVID-19 pandemic may have particular impact on assay sensitivity. Lack of adherence to investigational treatment regimens (failing to take study drug or failing to use the study device) can make the two regimens appear similar, since neither group is receiving any treatment (Food and Drug Administration 2016). Obtaining efficacy assessments off schedule can also affect conclusions if the outcome is dependent on timing, such as the relative time between receiving the investigational treatment and the assessment. Missing data will also affect interpretation, since probability of missingness might be related to the unobserved outcomes. Like missing data, it is risky to assume that other lack of rigor occurs independently of the outcome, and in general ignoring every subject who has deviations from the study plan (such as is done in a per-protocol analysis) is not appropriate.

Although superiority trials should also be conducted with rigor, the implications of lack of rigor in the context of NI studies are in some ways reversed. To the extent that lack of compliance with study treatment administration, timing of efficacy assessments and incomplete data make the two treatments appear more similar to each other, lack of rigor can reduce the power in superiority trials but increase the Type I error rate in noninferiority trials.

Because lack of rigor may increase the false positive rate of noninferiority trials, it is vital to not only have rigorous study conduct but also to demonstrate it. Without clear documentation, conclusions from even the most rigorously conducted trial will be questioned. It is therefore important to document ways in which rigor was enforced and reasons for failing to meticulously follow the study protocol, including issues specific to the pandemic.

2.2. How to Maintain Assay Sensitivity During COVID-19

Maintaining rigor in light of external events such as the COVID-19 pandemic requires adherence to some key principles. In this section, we outline some particular ways in which ongoing noninferiority studies can be operated to maximize the information and minimize bias in the conclusions.

2.2.1. Keep Subjects in Study and Collect Data

The first principle that we propose is that all subjects should be maintained in the study, to the degree possible, even in the event of external issues such as the COVID-19 pandemic. Subjects should continue to receive treatment and assessments should be obtained as close as possible to the scheduled timelines. Following subjects for safety assessments is important to fully understand the risks of an investigational treatment, including risks of stopping treatment. Additional information may be gained by understanding the impact of restarting treatment after an unexpected treatment interruption, although this is obviously not an objective of the study as stated when the study began. Finally, if subjects are benefiting from treatment, there is an ethical consideration to allow them to continue receiving that treatment.

Obviously, safety of study subjects is paramount. If remaining in the study poses a risk to subjects, discontinuing from the study may be in the best interest of the subject. If leaving home to receive treatment or provide efficacy assessments leads to an unacceptable risk in exposure to infectious pathogens, subjects may prefer to discontinue from the study. If stopping and restarting study treatment poses a risk, then permanently discontinuing treatment may be in the best interest of the study subject, while maintaining other study procedures if the subject agrees. While site personnel should counsel subjects on the advantages of remaining in the study in each situation, a subject who decides that the risk of continuing is not worth the benefit will be allowed to discontinue.

Often there are questions from site personnel, study subjects or even colleagues about the reason to collect data when subjects are not fully compliant with study procedures, especially when subjects discontinued investigational treatment (Little et al. 2012). While data obtained after treatment interruption due to the COVID-19 pandemic will often be of limited value in a noninferiority, we advocate collecting it to serve as supporting information and, importantly, to serve as primary information in the event that documented reasons for missing end up not fully supporting a conclusion that missingness was caused by the COVID-19 pandemic. Importantly, the collected data should be relevant for the estimand of interest. Arguably, for patients who discontinued assigned treatment due to pandemic related reasons, we may be interested in outcomes that would have been observed under normal circumstances, leading to hypothetical rather than treatment policy strategy. This is discussed in detail in Section 3.

2.2.2. Maintain Standards

Also of importance is the need to maintain standards for data collection, completeness, and quality. Data should be collected consistently and completely, as specified by the study protocol. While many of these recommendations are the same for superiority and noninferiority trials, we emphasize them for noninferiority trials because, as noted above, lack of (documented) rigor may reduce the (perceived) assay sensitivity of a study.

In any clinical trial, there may be pressure to reduce the burden on study subjects or sites. Particularly for a subject who discontinues study treatment in any study, there may be a desire on the part of the subject to reduce the frequency or intensity of data collection. When this happens, and site personnel insist on obtaining every piece of information at every visit, the subject may decide to fully withdraw from the study so that no further data are collected. An alternative is to focus data collection on the most important data points: first primary endpoints, at primary timepoints; then secondary efficacy endpoints, particularly at the most important timepoints, and primary endpoints at other timepoints; and so on. Prioritizing data collection in this manner may not result in full data being available, but may result in the most important data being available.

Collecting some data by phone, mail or internet is difficult to support. While this could work for select endpoints such as self-reported adverse events or patient rating scales, the scales must be validated for such collection methods. However, it is known that answers obtained through home administration may differ from answers obtained through office administration (Fairclough 2002).
3. Intercurrent Events and Missing Data

With all efforts described above, there may still be an important amount of missingness due to COVID-19 and related issues. In this section, we discuss some issues around analysis when data are missing.

First the estimand must be considered. As described in ICH E9(R1) (International Conference on Harmonisation 2019), the estimand is the precisely defined treatment effect in the population corresponding to the clinical question of interest. The precise definition includes clearly prespecified consideration of the treatment regimens compared, handling of each relevant intercurrent event (ICEs; events that may lead to deviation from the treatment regimens of interest or be part of them), the population in which the treatments are compared and the endpoint of interest and population level summary of treatment effect. For detailed discussion of application of ICH E9(R1) guidelines in clinical practice, see Mallinckrodt et al. (2020) and references therein. During the COVID-19 pandemic, specific ICEs exist that may interfere with the assigned treatment regimen, complicate collection and interpretation of outcomes and result in missing data due to effects of the pandemic. Typically, an ICE caused by COVID-19 may be classified into one of the following categories:

1. Discontinuation or modification of assigned treatment regimen and/or lack of assessments due to COVID-19 infection.
2. Discontinuation of assigned treatment regimen and/or lack of assessments due to lack of availability of treatment (drug supply, infusion procedure, etc.) or extra challenges in obtaining it that motivate discontinuation.
3. Lack of outcome assessment or substitution of standard assessment with a different procedure (e.g., via telephone or internet) making interpretation of outcome difficult (even if subjects may remain on treatment). This last type arguably may be considered not an ICE, as no change in treatment regimen actually occurs but rather change in the ability to collect and interpret the outcome data.

As a result of these ICEs, missing data can arise.

To better understand the impact of missing data caused by COVID-19 on analysis, it is useful to review the classification of missingness mechanisms. Missingness is categorized according to the relationship between the patient-level data and probability of missing outcomes. The data here include outcomes (both observed and unobserved due to missingness), assigned treatment and patient baseline covariates. When missingness is unrelated to observed or unobserved data, the missing data are called missing completely at random (MCAR). When missingness depends on observed data but, conditionally on observed data, is unrelated to unobserved data, the missing data are called missing at random (MAR). Otherwise, the missing data are called missing not at random (MNAR). Under MCAR or MAR valid inference about parameters governing outcome process can be based on modeling observed outcomes ignoring the parameters underlying the missingness process (provided the regularity condition of parameter separability is satisfied). Specifically, likelihood and Bayesian inferences are ignorable under MAR and MCAR whereas nonlikelihood-based inferences (e.g., based on generalized estimating equations, GEE) are ignorable only under MCAR (see Molenberghs and Kenward 2007). Data that are MNAR do not lend to ignorable analysis, since the analysis must jointly model outcome and missingness process to ensure valid conclusions. As joint modeling of outcome and the missingness process under MNAR is very challenging, often analysis proceeds assuming MAR. It is important to emphasize that it is not possible to test using only observed data whether the data are MAR or MNAR; therefore any analysis requiring MAR is based on a strong and untestable assumption.

We now consider typical situations when missing data are caused by COVID-19.

A site may be unable to fulfill its obligations under the protocol; site personnel may be called to treat infected patients in a hospital setting or triage suspected infections in an emergency setting, personal protective equipment may be unavailable to site personnel during protocol-required invasive procedures, study subjects may be prohibited from traveling to study sites due to shelter-in-place orders, study subjects may be hesitant to travel outside of the home due to the risk of infection, or study subjects may have suspected or confirmed cases of COVID-19 infections. Many but not all of these events can be almost certainly ascribed to causes other than the treatment under investigation and are independent of subject-level outcomes or characteristics, and therefore the resulting data missing due to those causes is MCAR. Other of these events—notably, discontinuation due to confirmed infection with COVID-19—can never be fully ascribed to causes other than the treatment of interest, and resulting missing data should be handled as MAR or MNAR. But, at least at a superficial level, any missingness due to public response to the pandemic and not to treatment, adverse events or outcomes can be reasonably treated as MCAR.

We now further consider cases when the missingness process caused by COVID-19 may deviate from MCAR. A difference may be drawn between missingness that involves an individual subject’s decision and missingness that does not, which is clarified in the following. A subject who is receiving great benefit from an investigational treatment may be highly motivated to venture to a study procedure even during a pandemic while another who perceives less benefit, or even harm from an adverse event, may be highly motivated to shelter in place and skip a study procedure, especially if the subject has known risk factors that exacerbate the risk of COVID-19 complications. We recognize this requires careful assessment and documentation at study sites. Further, if a site closes and all subjects at that site discontinue treatment and assessments, site (an observed baseline covariate) is a predictor of missingness. Thus, even in this situation, data that are missing for reasons clearly independent of any observed or unobserved outcome data might not meet the definition of MCAR and need to be considered MAR for appropriate analysis.

Subsequent discussion in this article will focus on addressing data that is missing due to the impact of the pandemic with the “hypothetical strategy” estimand, in which the value that would have been observed, had the COVID-19 pandemic not prevented it from being observed, is of interest. This estimand will ignore any data collected after treatment interruption due to site closures or shelter-in-place orders due to the pandemic, and will treat missing data due to these reasons as MCAR or as MAR.  [Notably, as recognized by Meyer et al. (2020), only
ICEs due to the immediate impact of the pandemic are cause for ignoring subsequent data under this paradigm; other ICEs should be handled according to the strategy specified when the study was started, to address the estimand of interest. While a “treatment policy strategy” might be of value for supportive or sensitivity analyses of superiority studies, we do not see value for noninferiority studies. The treatment policy strategy generally uses all data collected, in spite of ICEs, in the analysis. If a superiority study using a treatment policy strategy approach to ICEs shows an advantage of an investigational treatment in spite of the treatment discontinuations associated with the COVID-19 pandemic, it may provide additional evidence of benefit of the treatment because a treatment effect was apparent in spite of the difficulties imposed by the ICEs. However, even in the case of superiority trial, “treatment policy” may be of very limited value as having questionable generalizability to the “after pandemic” situation (see, e.g., Mallinckrodt et al. 2020, chap. 4). A noninferiority analysis that uses such data may be biased toward the alternative hypothesis, since equality of outcomes is not part of the null, and will therefore be of limited value in a rather unique situation such as the COVID-19 pandemic.

### 3.1. Impact on Study Conclusions

For COVID-19 and related reasons, the reason for missing data must be carefully documented. Unfortunately, it is not always possible to gather such information with complete confidence. A subject who discontinues the study is unable to give any further information, and sites are legally prohibited from additional probing of a subject who has withdrawn consent. A site that is closed will not be able to gather information on whether a subject would have been available for an evaluation, had the site been open. The onus will be on the sponsor to ensure adequate documentation of the reason for missing data, and without clear documentation, it may be required to assume that missing data are not MCAR leading to sensitivity analyses under non-MCAR missingness. Thus, it is vital to carefully document reasons for missing data or delayed treatment; if documentation is determined post hoc to be inadequate, using the hypothetical strategy estimand and assessment as MCAR or even MAR may not be defensible.

When data are MCAR, an analysis that ignores missing data can provide valid conclusions; however, we do not recommend it due to inefficiency. A simple application of this principle is to ignore any subject with data missing for reasons unrelated to any patient-level data. With no data missing for any other reason, such an analysis would simply exclude any subject with data that are MCAR. While the results of the analysis would be valid, such an analysis would also be inefficient since any data obtained from subjects without complete data would be excluded, while using that data could provide additional information. The goal, then, is to use a method that includes subjects with incomplete data until the time of a COVID-19 discontinuation of treatment.

When analyzing a dataset with incomplete longitudinal data, three general categories of analysis methods are available: direct likelihood, multiple imputation, and various nonlikelihood based methods enhanced with adjustments for missing data (such as weighted GEE). Both direct likelihood and multiple imputation are likelihood-based often extending in Bayesian versions, especially multiple imputation that typically uses sampling from predictive posterior distributions (see Mallinckrodt and Lipkovich 2016 and references therein). Direct likelihood methods typically start by posing likelihood for complete longitudinal data (e.g., multivariate normal for continuous case, conditional on baseline covariates) and then proceed to observed data likelihood by integrating out missing data. Under ignorability (requiring MAR and additional separability condition for parameters governing outcome and missingness processes) the maximal likelihood solution from observed data is consistent for the parameters of the underlying complete data likelihood, provided of course that the analysis model is correctly specified. Note that direct ML methods are challenging for modeling longitudinal binary data which is the focus of this article (with some discussion of other models in Section 4). Therefore, we will consider multiple imputation strategies (as will be illustrated in simulations later in this article). Multiple imputation (in the context of longitudinal data) models the relationship among longitudinal outcomes and baseline covariates and uses an estimated model to impute missing values given observed values. By employing Bayesian methods for constructing imputation models, imputing missing data amounts to sampling from posterior predictive distributions of missing data given observed data. Doing this multiple times, not only can the treatment effect be estimated, but the uncertainty due to missing data can be properly assessed. A popular approach is based on Rubin’s (1987) combination rules: first M imputed sets are constructed, then estimated treatment effects and associated standard errors from each individual set are combined. A point estimate is computed simply as the arithmetic mean over individual estimates and a valid (although often conservative) confidence interval is constructed by using Rubin’s rules for total variance incorporating both within and between imputation variability. While both direct likelihood and MI-based methods are more complex than simply ignoring any data that is missing, both allow for subjects with partial data to provide information for the analysis, resulting in more efficient analysis while avoiding making strong assumptions on data being MCAR.

While single imputation strategies are typically not recommended, worst case analysis may be considered when the primary outcome is binary. In a disease that has no effective or approved treatment and does not resolve spontaneously, it is unlikely that a subject will discontinue the study if the subject notices efficacy and does not experience a safety or tolerability issue. In such a study, a subject who discontinues or otherwise does not have the primary outcome observed can reasonably be assumed to have had a suboptimal outcome, and imputed as a nonresponder (nonresponse imputation, or NRI). NRI is problematic with missing data that are known to be MCAR—in fact, NRI is predicated on knowing that missingness is related to suboptimal (unobserved) outcome, and therefore implicitly assumes data are not MCAR. Another way of looking at NRI imputation is considering it an implementation of composite strategy for handling ICEs, which is especially easy for binary outcomes, when the ICE becomes part of the defined outcome. In this way, the missing data simply does not exist, as
we consider that for a patient with ICE the observed outcome is always the worst case (nonresponse).

3.2. Simulation Methods

Meyer et al. (2020) recommended studying the impact of pandemic-related issues with simulation studies making specific assumptions about factors that may be relevant for missingness. In this section, we give an example of a small simulation study to evaluate the impact of missing data in the NI setting under various methods of handling missing outcomes relevant to COVID-19. As a typical example encountered in many NI studies we consider a binary outcome \( Z \), defined by dichotomizing changes from baseline in a continuous \( Y \) outcome, so that \( Z_t = I(Y_t < c) \) for all post-baseline visits, \( t = 1, \ldots, T \). The interest is in comparing the probability of clinical response \( (Z_1 = 1) \) for the experimental and control treatments at the last study endpoint. The NI null hypothesis is tested against \( H_0: \pi_0 - \pi_1 \geq \Delta \) versus \( H_1: \pi_0 - \pi_1 < \Delta \), where \( \pi_0, \pi_1 \) are probabilities of responder status at last visit for control and experimental treatment, respectively, and \( \Delta > 0 \) is a NI margin that quantifies acceptable deterioration in efficacy of a new experimental treatment to be considered noninferior to established control.

The simulations mimicked a clinical trial in which longitudinal continuous outcomes were observed and binary outcome (clinical response) of interest was defined by dichotomizing the continuous outcome. This setting can be contrasted with a situation when the endpoint of interest is naturally a binary variable that may be correlated with intermediate continuous outcomes (see Lipkovich and Wiens 2018). We simulated continuous outcomes using multivariate normal data, including baseline and three post-baseline values, evaluated at weeks 8, 16, and 24, then dichotomizing the last value into success and failure (response and nonresponse). Dichotomization was tuned to provide target success rates, \( \pi_0 = 0.5, \pi_1 = 0.375 \) under NI null, corresponding to the NI margin \( \Delta = 0.125 \). A sample size of 250/arm was used to provide 80% power to reject the null hypothesis when the two treatment arms have identical response rates of 50%, assuming noninferiority margin of 0.125, and no loss of data due to dropouts. The NI hypothesis was evaluated for complete data by fitting a logistic regression including terms for baseline continuous outcome \( Y_0 \) and treatment indicator \( R = 0, 1 \). Then the NI null was rejected if the upper 95% confidence limit for the difference in proportions is below the NI margin, that is, \( \hat{\pi}_0 - \hat{\pi}_1 + 1.96 \times SE < \Delta \), where \( \hat{\pi}_r = n^{-1} \sum_{i=1}^n \hat{\pi}_{ri} \) are the marginal probabilities for treatments \( r = 0, 1 \) computed by averaging individual probabilities \( \hat{\pi}_{ri} = 1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 r - \hat{\beta}_2 Y_{0i})^{-1} \) predicted from logistic model for patients pooled from both treatment groups \( n = 500 \); the standard errors \( SE = \sqrt{\text{var}(\hat{\pi}_0 - \hat{\pi}_1)} \) were computed using delta method (Ge et al. 2011; Qu and Luo 2015) with bias-adjustment to account for randomness in baseline covariates (Bartlett 2018).

To thusly generated complete data, a monotone missingness process was applied (i.e., missing values due to dropouts only) resulting in expected proportions of missing outcomes at the last scheduled visit (10%, 15%, or 20%). We considered scenarios in which missingness was MCAR (mimicking broad-based shelter-in-place orders, as an example) and scenarios in which missingness was MAR (depending on the subject's choice, possibly based on perceived lack of benefit compared to the risk associated with traveling to the visit). For MCAR data, we applied at each evaluation visit a fixed probability of dropouts so as to ensure the desired probability of missing outcomes at study end. For MAR data, the missingness process was such that at each visit starting from second post-baseline, probability of dropout depended on current and previous continuous outcomes, so that subjects who had better early responses at the current evaluation visit \( t \) and previous visit \( t - 1 \) were less likely to have missing data at the next timepoint, \( t + 1 \). Specifically, the logit of probability of discontinuing study at visit \( t + 1 \) was modeled as \( \logit(P(Y_{t+1} = 1)) = a_0 + a_1 U \), where \( U = I(Y_t > c_1 \& Y_{t-1} > c_2) \) or \( Y_t \leq c_1 \& Y_{t-1} > c_3) \). The cutoffs \( c_1, c_2, c_3 \) were chosen to quantify poor early response (either large level of current \( Y \) with modest or no improvement from previous visit, or very large worsening from previous evaluation) and coefficients \( a_0, a_1 \) where calibrated to ensure the desired strength of the relationship between probability of dropout and lack of early response (\( U \)) and the overall dropout rate in the control arm (at study end). Note that when the null hypothesis is true, dropout rates in the investigational group will be higher than in the control arm, a difference naturally caused by worse intermediate outcomes in the investigational arm. A small additional probability of dropout at the first post-baseline visit was applied.

Six methods of handling missing data were considered:

- Observed case, in which only subjects with observed data at the last timepoint were included in the analysis
- NRI, in which subjects with missing data at the last timepoint were imputed as nonresponders (treatment failures)
- Four strategies that used multiple imputation, differing by variables in the imputation model and data used to estimate imputation models:

  - Strategy 1: Sequential imputation strategy for monotone missingness with Bayesian regression (Schafer 1997) was used to impute missing continuous outcomes \( Y_t, t = 1, 2, 3 \) (separately by treatment arms). Specifically, subsequent values were imputed using a Bayesian linear regression with baseline score and responses up to and including the current timepoint; the imputed continuous values at the last timepoint are then dichotomized to produce binary outcome, \( Z_3 \).

  - Strategy 2: Continuous data observed at baseline and at timepoints prior to discontinuation was used to impute a continuous value up to the next-to-last timepoint, as in strategy 1, then Bayesian logistic regression was used to impute the binary value at the last timepoint using baseline and earlier continuous outcomes \( Y_t, t = 0, 1, 2 \) as predictors (separately by the two treatment arms).

  - Strategy 3: Similar to strategy 2, but with only the baseline (prerandomization) value \( Y_0 \) used as predictor of the logistic model (separately by the two treatment arms).

  - Strategy 4: Similar as strategy 2, but with imputation for both arms done using the imputation models estimated on the control arm alone; the noninferiority margin (\( \Delta \)) sub-
tracted from individual probabilities (drawn from posterior distribution estimated from control arm) for subjects in the investigational arm, to mimic the NI null case.

For all imputation strategies, 50 imputations were used, and the estimated differences in proportions and associated standard errors (computed using delta method) were combined using Rubin’s rules and NI null was rejected if the upper bound of a one-sided 97.5% CI for the difference was below the noninferiority margin.

In the following, we provide brief rationale for each of the imputation strategies.

**Strategy 1.** has been widely used and recommended as a standard for analyzing binary endpoints derived based on continuous outcomes (Lipkovich, Duan, and Ahmed 2005; Mallinckrodt and Lipkovich 2016). It is attractive because it uses all available continuous data for imputation so loss of information when dichotomizing imputed continuous response occurs only at the very last point.

**Strategy 2.** replaces dichotomization at the last step with fitting a Bayesian logistic model. The binary endpoint for patients with missing endpoint is simulated as Bernoulli random variables with the logits of individual probabilities computed as linear predictor in $Y_0$, $Y_1$, $Y_2$ with regression coefficients drawn from their posteriors. The flexibility of imputation based on the logistic model allows for various sensitivity analyses (as in strategy 4). While using imputation on the binary scale may cause some loss of power compared with imputation on continuous scale up to last timepoint following by dichotomization (strategy 1), it is not clear whether this is actually the case for a given set of scenarios, or the magnitude of the loss; therefore assessing operating characterize of this easy to implement and attractive for practitioners strategy in simulation is important.

**Strategy 3.** ignores all intermediate outcomes $Y_1$, $Y_2$ when imputing final outcome. One can argue that this strategy should be valid when data are known to be MCAR, however, it is not clear whether, or to what degree, using intermediate outcomes may improve power under MCAR.

**Strategy 4.** is a version of the copy reference approach, combined with a delta adjustment to provide imputations under the NI null hypothesis. Here, it illustrates the flexibility of multiple imputation in general that allows one to manipulate with parameters of the imputation model before simulating imputed values from predictive posterior distributions. In the context of imputing binary outcome for a NI study, we subtracted $\Delta$ from each individual probability of response in subjects on the investigational treatment (or imputed 0 if the original value was less than delta). As a result, imputed values were mimicking those expected if NI was indeed true for all subjects who discontinued treatment, $\pi_{1i} = \pi_{0i} - \Delta$.

### 3.3. Simulation Results

As a benchmark, complete data (before missingness was applied) were analyzed, to confirm Type I error rates, power, and bias.

Estimates of bias from various imputation methods were consistent with results expected after the Type I error rate and power were assessed, and so are not reported in this article. Methods that led to inflated Type I error rate under the null hypothesis also led to underestimate of the treatment difference; methods that led to reduced power under the alternative also led to overestimate of the treatment difference.

#### 3.3.1. Type I Error Rate

Type I error rates from the simulations are shown in Table 1. Each value is based on 10,000 simulated studies, with a target of $\alpha = 0.025$ (for the one-tailed test). If a method controlled the Type I error rate at 2.5%, the observed Type I error rate will be under 2.75%, 95% of the time (assuming normal approximation for the rejection rate with mean at $\alpha$ and variance at $\alpha(1 - \alpha)/10,000$.)

When data were MCAR, all methods except NRI controlled the Type I error rate. The reason why NRI produces inflation in Type I error even under “seemingly harmless” MCAR is easy to understand if we recall that under MCAR the dropout rate $(d)$ is the same across treatment arms, therefore the expected proportions under NRI is $\pi_0 d + (1 - d) \times 0 = \pi_0 d$ for control and $\pi_1 d$ for treated, and the treatment difference $(\pi_1 - \pi_0) d = \Delta \cdot d$. Therefore, the effect shrinks proportionally to the amount of missing data with the expected bias $= \Delta \cdot (d - 1)$.

Multiple imputation often provides slightly conservative inference due to using Rubin’s formula for variance. Partly it is due to the finite number of imputation (which for analysis of actual data is advised to be made sufficiently large but was held at $m = 50$ in this simulations), partly because of inconsistency of Rubin’s variance estimator (see discussion in Wang and Robins (1998) and Robins and Wang (2000)). The most conservative of imputation strategies was strategy 1 with a Type I error rate of 0.017. (The ratio of model-based to the true (simulation) standard error was within 3%–7%, results not reported here.)

When data were MAR, again all methods except NRI controlled the Type I error rate. NRI produced a Type I error rate that was less biased under MAR than under MCAR (0.048 with 20% of data MAR compared to 0.086 with 20% of the data MCAR). Multiple imputation with strategy 1 or with strategy 4 was the most conservative with a Type I error rate of around 0.02.

Judging by the simulation results of Type I error alone, all proposed methods except NRI appear acceptable when data are assured to be MCAR or MAR.

| Dropout rate | 10% | 15% | 20% | 10% | 15% | 20% |
|--------------|-----|-----|-----|-----|-----|-----|
| Complete data | 0.0256 | 0.0256 | 0.0256 | 0.0256 | 0.0256 | 0.0256 |
| Strategy 1 | 0.0215 | 0.0216 | 0.0169 | 0.0214 | 0.0201 | 0.0198 |
| Strategy 2 | 0.0263 | 0.0265 | 0.0258 | 0.0258 | 0.0263 | 0.0274 |
| Strategy 3 | 0.0256 | 0.0257 | 0.0244 | 0.0263 | 0.0277 | 0.0281 |
| Strategy 4 | 0.0240 | 0.0243 | 0.0215 | 0.0197 | 0.0208 | 0.0199 |
| NRI | 0.0499 | 0.0673 | 0.0862 | 0.0357 | 0.0414 | 0.0475 |
| Observed data | 0.0255 | 0.0259 | 0.0244 | 0.0251 | 0.0273 | 0.0290 |
Table 2. Estimated power using six different analysis strategies when 10%, 15%, or 20% of the outcome data is missing at study endpoint, either MCAR or MAR.

| Dropout rate | MCAR       | MAR       |
|--------------|------------|-----------|
|              | 10%        | 15%       | 20%        |
| Complete data| 0.812      | 0.812     | 0.812      |
| Strategy 1   | 0.782      | 0.769     | 0.755      |
| Strategy 2   | 0.772      | 0.753     | 0.734      |
| Strategy 3   | 0.767      | 0.745     | 0.717      |
| Strategy 4   | 0.731      | 0.661     | 0.573      |
| NRI          | 0.815      | 0.816     | 0.825      |
| Observed data| 0.770      | 0.746     | 0.721      |

3.3.2. Power

Power from the simulations is shown in Table 2. Each value is based on 5000 simulated studies, with a target of 80% from the test with no missing data. If a method has 80% power, the observed power will be between 78.9% and 81.1% approximately 95% of the time.

When data were MCAR, all methods except NRI imputation controlled the Type I error rate and can be considered. Multiple imputation with imputation strategy 1 had the highest observed power at over 75%, followed by strategy 2 with over 73% power, with 20% missing data. Each imputation strategy showed a reduction in power compared to the situation in which no data are missing, but an improvement compared to the power (nearly 72%) that was seen using only observed data. Thus, missing data causes a loss of up to 8 percentage points in power, and multiple imputation adds back 2–4 percentage points of power by using information from subjects who have some, but not complete, data available, in the circumstances studied.

When data were MAR, multiple imputation using imputation strategy 1 again had the highest observed power, with over 74% power when 20% of the data were missing.

Imputation strategy 4 has the lowest power by design, as it intends to consider the worst case, assuming outcomes for patients who discontinue follows the null.

4. Summary and Conclusions

We consider analysis of noninferiority trials in which some data are missing due to the impact of the COVID-19 pandemic. Some of the advice provided in this article differs from advice provided for studies run before or after the current pandemic. The justification for the differing advice is that the immediate impacts of the COVID-19 pandemic, including large-scale shelter-in-place requirements and acute shortages of personal protective equipment, are not expected to be long-lasting or repeated. This logic supports the use of the hypothetical strategy estimand with analyses that are robust to data MAR.

When a site is closed due to the pandemic and unable to treat or assess study subjects, some data will be clearly MCAR. Analysis of observed data will control the Type I error rate, but incorporating partial information from subjects without complete data will improve the power.

Multiple imputation methods are only as good as the model used to impute missing data. Our simulations indicated that multiple imputation with imputation strategy 1, imputing values of the underlying continuous variable through the final timepoint and dichotomizing the imputed value to assess response, was the most powerful of the methods that controlled Type I error rate and bias. Other methods either produced inflated Type I error rates, lower power or biased point estimates of treatment effect.

In this simulation study, we were interested in evaluating NI hypothesis for the difference in marginal means (averaged probabilities of responders over the target population), and therefore conducted inference using methodology for marginal probabilities (see Bartlett 2018). Arguably, one could posit a NI margin for the difference in conditional probabilities evaluated at specific values of baseline covariates; however, the generalizability of such estimand may be of limited value. Conversely, the estimator for marginal probabilities generalizes to future data, assuming the same target population (in terms of baseline covariates), however, like analysis based on conditional estimand, it may have questionable applicability for populations with baseline covariates very different than those assumed in a clinical trial.

When developing an imputation model that uses preliminary timepoint data to impute final timepoint data in a noninferiority trial, it will be important to understand the response patterns in the two arms. Treatments that produce identical response rates at the final timepoint may not produce identical patterns at preliminary timepoints. Especially, when the mechanism of action is different for the two treatments, it is important to consider imputation models that use only data from one treatment arm to impute missing data from that treatment arm, to avoid inflating the Type I error by making the two arms look more similar than they are. We close by noting that this is the opposite from the situation in superiority trials, demonstrating again that noninferiority trials are different from superiority trials.

Wiens and Rosenkranz (2013) assessed various methods of addressing missing data in the context of continuous outcome when a direct likelihood approach (specifically, MMRM) was an appropriate analysis method. They concluded that, among methods that controlled the Type I error rate, direct likelihood was the best choice due to high power and low bias. Specifically, when missing data were MCAR, MMRM was shown to have higher power than an analysis that ignored subjects with missing data. Again, this demonstrates that incorporating data obtained before the MCAR-causing event will increase power compared to completely ignoring subjects with data that are MCAR. We did not consider direct likelihood approach for analysis of binary outcomes (see Breslow and Clayton 1993) due to limitations on the model, in particular modeling serial correlations (important for analysis of clinical trial data) requires doubly iterative procedures based on linearized pseudo-likelihood (e.g., implemented in SAS proc GLIMMIX), and the potential for biased results in certain situations (Lipkovich, Duan, and Ahmed 2005). More importantly for our setting, unlike MI-based procedures direct likelihood for binary data would not allow us to take advantage of the information contained in the continuous scores underlying binary outcomes.

An equivalence study aims to demonstrate that an investigational treatment is neither much worse nor much better than an active control. To the extent that the requirements for an equivalence study are the same as for a noninferiority study (need for assay sensitivity, in particular), the implications for an
equivalence study conducted during the COVID-19 pandemic are the same as for a noninferiority study.

In summary, the conduct of noninferiority trials during the COVID-19 pandemic requires careful planning and implementation. Collecting data, including reasons for missing data, is important to support inference. While analyses that ignore data that is MCAR provide valid inference under strong assumptions about the missing data mechanism, it is unwise to do so in situations we studied because, first, analysis including partial data are more powerful and, second, any doubts about whether the data are MAR instead of MCAR are moot when analysis methods appropriate for MAR data are used.

Acknowledgments

The authors thank Dr. Yongming Qu for review of a draft article and helpful comments, and two anonymous referees for helpful reviews.

References

Bartlett, J. W. (2018), "Covariate Adjustment and Estimation of Mean Response in Randomised Trials," Pharmaceutical Statistics, 17, 648–666. [447, 449]

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," Journal of the American Statistical Association, 88, 9–25. [449]

Fairclough, D. (2002), Design and Analysis of Quality of Life Studies in Clinical Trials, Boca Raton, FL: Chapman & Hall/CRC Press. [444]

Food and Drug Administration (2016), Non-Inferiority Clinical Trials to Establish Effectiveness, Washington, DC: Food and Drug Administration. [443, 444]

Ge, M., Durham, L. K., Meyer, R. D., Xie, W., and Thomas, N. (2011), "Covariate-Adjusted Difference in Proportions From Clinical Trials Using Logistic Regression and Weighted Risk Differences," Drug Information Journal, 45, 481–493. [447]

International Conference on Harmonisation (1998), Statistical Principles for Clinical Trials E9, London: International Conference on Harmonisation. [443]

——— (2019), Addendum on Estimands and Sensitivity Analyses in Clinical Trials E9(R1), London: International Conference on Harmonisation. [445]

Molenberghs, G., and Kenward, M. G. (2007), Missing Data in Clinical Studies, Chichester: Wiley. [445]

Lipkovich, I., Duan, Y., and Ahmed, S. (2005), "Multiple Imputation Compared With Restricted Pseudo-likelihood and Generalized Estimating Equations for Analysis of Binary Repeated Measures in Clinical Studies," Pharmaceutical Statistics, 4, 267–285. [448, 449]

Lipkovich, I., and Wiens, B. L. (2018), "The Role of Multiple Imputation in Noninferiority Trials for Binary Outcomes," Statistics in Biopharmaceutical Research, 10, 57–69. [447]

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A. and Neaton, J. D. (2012), "The Prevention and Treatment of Missing Data in Clinical Trials," New England Journal of Medicine, 367, 1355–1360. [444]

Mallinckrodt, C., and Lipkovich, I. (2016), Analyzing Longitudinal Clinical Trial Data: A Practical Guide, Boca Raton, FL: Chapman & Hall/CRC Press. [446, 448]

Mallinckrodt, C., Molenberghs, G., Lipkovich, I., and Ratitch, B. (2020), Estimands, Estimators and Sensitivity Analysis in Clinical Trials, Chapman & Hall/CRC Biostatistics Series, Boca Raton, FL: Chapman & Hall/CRC Press. [445, 446]

Meyer, R. D., Ratitch, B., Wolbers, M., Marchenko, O., Quan, H., Li, D., Fletcher, C., Li, X., Wright, D., Shentu, Y., and Englert, S. (2020) "Statistical Issues and Recommendations for Clinical Trials Conducted During the COVID-19 Pandemic," Statistics in Biopharmaceutical Research (in press). [443, 445, 447]

Qu, Y., and Luo, J. (2015), "Estimation of Group Means When Adjusting for Covariates in Generalized Linear Models," Pharmaceutical Statistics, 14, 56–62. [447]

Rothmann, M. D., Wiens, B. L., and Chan, I. S. F. (2012), Design and Analysis of Non-Inferiority Trials, Boca Raton, FL: Chapman & Hall/CRC Press. [443]

Robins, J. M., and Wang, N. (2000), "Inference for Imputation Estimators," Biometrika, 87, 113–124. [448]

Rubin, D. B. (1987), Multiple Imputation for Nonresponse in Surveys, New York: Wiley. [446]

Schafer, J. L. (1997), Analysis of Incomplete Multivariate Data, New York: Chapman & Hall. [447]

Wang, N., and Robins, J. M. (1998), "Large-Sample Theory for Parametric Multiple Imputation Procedures," Biometrika, 85, 935–948. [448]

Wellek, S. (2010), Testing Statistical Hypotheses of Equivalence and Noninferiority, Boca Raton, FL: Chapman & Hall/CRC Press. [443]

Wiens, B. L., and Rosenkranz, G. K. (2013), "Missing Data in Noninferiority Trials," Statistics in Biopharmaceutical Research, 5, 383–393. [449]