Missing Data in Non-Inferiority Clinical Trials

by

Brooke Ann Rabe

A Dissertation Presented to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN STATISTICS

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

In the Graduate College

The University of Arizona

2019

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Brooke A Rabe titled Missing Data in Non-Inferiority Clinical Trials and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Melanie L Bell, PhD — Date: 5/7/19

Edward J Bedrick, PhD — Date: 5/7/19

Chiu-Hsieh (Paul) Hsu, PhD — Date: 5/7/19

Joseph C Watkins, PhD — Date: 5/7/19

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Melanie L Bell, PhD — Date: 5/17/19
Dissertation Committee Chair
Epidemiology and Biostatistics Department

Acknowledgements

First, I thank my PhD supervisor, Dr. Melanie Bell, without whom these contributions could not have been possible. She might have preferred to be climbing or biking; still, she devoted many hours to discussing ideas, deciphering my prose, and helping me manage time. For her scholarship and guidance, financial support, and encouragement, I am grateful forever.

Sincere thanks to Dr. Edward Bedrick, Dr. Paul Hsu, and Dr. Joe Watkins, for offering their time as members of my dissertation committee. I thank them for their availability throughout my research journey and for their helpful comments on my dissertation which ultimately improved the quality of this work. Dr. Watkins, who as the Chair of the Statistics GIDP, was instrumental in my success as a graduate student, years ago in 2009 kindled my interest in Probability and Statistics. I am so incredibly lucky to have had Joe in my corner.

I am extremely grateful for my time at Cockins Hall with The Ohio State University Statistics Department. Thanks to Drs. Elizabeth Stasny, Noel Cressie, and Kate Calder who each influenced and deepened my understanding of Statistics. Thanks also to Dr. Prem Goel and Dr. Mark McCord in Civil Engineering at OSU who supported me as a Graduate Research Assistant.

Thanks to the professors at the University of Arizona and Pima Community College who unleashed my excitement about Math and Physics: Tony Pitucco, Ana Mantilla, Ted Laetsch, Rob Indik, Bill Velez, Jason Hsu, and Feryal Őzel. Thanks to Professor Mike Evans (now at University of Maryland) for introducing me to big data and modeling.

To the many friends, labmates, collaborators and mentors who enriched my education: Libby Floden, Mary Miller, David Garcia, Uma Nair, Tracy Crane, Suz Tolwinski-Ward, Kevin Anchukaitis, James Little, Michael Sohn, Rex Hu, Dan Luo, Meng Lu, Sara Zeibell, Dailu Chen, Kyle Carter, Nick Lyttal, Rhoda Muse, Kira Gonzalez, Kevin Keys, Jonah Beaumont, Emile Doering and Lila Sideras.

To the Mathematics Department and the Center for Health Disparities Research for their financial support. Thanks to Lingling An for her generous support and giving me a chance. To my MATH 263 mentor, Kerima Ratnayaka, and to Tina Deemer for believing

in my teaching potential. To university staff: Melanie Bowman, Brooke Valmont, Cheryl Ekstrom, thank you so much for your support over the years.

I thank my mom, Meryl Midler, for her sensible, rationale mind, her love and generosity, and my dad, Gregg Rabe, for opening up my world to History, Literature and Art, and giving me a place to feel unconditionally accepted. Thanks to my late grandfather Rudy, no stranger to adventure, who taught me to aim high, and to the withered ancients for bringing me back to Earth. I love my wonderful Tucson family, the Koses, and thank them for their warmth, love, and humor.

Finally, thanks to my husband, Thomas. As I typeset homework on weekends, or agonized over timelines, he supported me in every way, holding my hand though the peaks and valleys from beginning to end.

Dedication

For Finn, Dashiell

and Tom, my own Mister Fahrenheit

Aren't we having a ball?

Contents

List of Tables

List of Figures

Abstract

In drug, device and behavioral clinical trials, patient withdrawal, loss-to-follow-up, and non-compliance with treatment protocols complicate analysis. When the data planned for collection are compromised or incomplete, estimates for treatment effect may be biased and trial conclusions may not be generalizable. Non-inferiority trials aim to show that an experimental treatment is therapeutically no worse than existing treatments. If a new treatment may be preferred for reasons such as lower cost, convenience, or improved safety profile, a non-inferiority design may be ideal for investigating whether the treatment is as efficacious as an active control within some pre-defined margin. Non-inferiority trials are by nature less conservative than superiority and placebo-controlled studies, and many of the challenges in their analysis and interpretation are exacerbated by missing data. Although missing data problems have been extensively studied, there has been little research on best practices for their handling in non-inferiority hypothesis testing to ensure control of type I error. I present a systematic review of non-inferiority trials to demonstrate the prevalence of missing data and understand current practices. Next, I conduct simulation studies to characterize the effects of missing data in non-inferiority trials. Using information from my systematic review of non-inferiority trials, I select methods that are commonly used as well as state-of-the-art, statistically-based methods under various missing data mechanisms. Finally, I develop a sensitivity analysis using a pattern-mixture model approach for multiple imputation adapted for the non-inferiority design. These are necessary for examining how sensitive the trial conclusions are to assumptions about missing data. Given the increasing popularity of the non-inferiority design, the persistent challenge of missing data and patient compliance, and the reliance of regulators and clinicians on trial results, there is a critical need to improve rigor and reproducibility of non-inferiority analyses. Better practices have potential for patients' easier access to new treatments and for minimizing risk of exposure to treatments that are ineffective.

Chapter 1

Introduction

1.1 Motivation

The randomized controlled trial (RCT) with a placebo control is perhaps the strongest tool for showing efficacy of a new treatment. Randomization balances the effects of unmeasured variables between treatment arms, and the use of placebo enables direct estimation of the treatment effect. However, in some cases, a placebo-controlled trial may be impractical or unethical. For example, if an effective treatment is available and delaying its administration puts patients at risk of disease progression or death, then a placebo assignment is not possible.

Non-inferiority trials aim to show a novel treatment is at least as effective as the standard of care. The design is an attractive choice when a new treatment is not expected to be superior in efficacy but has benefits such as improved safety, fewer side effects, lower cost, convenience of delivery, and so on[2]. Typically, the standard of care serves as the only control; therefore, it must have been shown to be effective in previous placebo-controlled RCTs.

Non-inferiority trials are more challenging to design, analyze and interpret than placebo-controlled or superiority studies. Defects in the design or deviations from trial protocol can result in the treatment arms appearing more similar, thereby skewing results in favor of the alternative conclusion of non-inferiority. Thus, the trials are by nature less conservative than superiority studies. Protocol deviations that are especially troublesome in non-inferiority

trials are patient dropout and the resultant missing observations. As in all experiments, unplanned missing data may lead to biased estimates and often reduce statistical power to detect clinically meaningful effects. However, missing data present additional challenges in the non-inferiority setting.

This dissertation was motivated by the broad challenges of missing data in non-inferiority RCTs and lack of guidance on statistically principled approaches to their handling. The non-inferiority design is increasing in popularity, possibly due to a combination of factors including the availability of many treatment options, fewer investments in new drug discovery, and the greater difficulty of proving superiority. As greater numbers of researchers, regulators and patients rely on treatment regimens that have been accepted based on evidence from non-inferiority trials, there is critical need to ensure robust conclusions. In the chapters that follow, I investigate the main sources of error due to missing data and other protocol deviations in non-inferiority conclusions and aim to improve statistical practice for their analysis.

1.1.1 Missing Data

Performing a correct statistical analysis of a designed experiment is more complicated when data that were planned for collection are missing. Observations may be missing for many different reasons. Sometimes, the reasons can be understood or inferred from external circumstances or expert knowledge. Other times, the causes of missing data are unknown and further assumptions need be made about the missing values. Regardless of whether these assumptions are correct, they should be communicated along with the results from analysis.

Although data can be missing from variables that are not of primary interest, such as demographic or baseline clinical covariates, this work focuses on measurements for the primary outcome, whether collected at a single time or repeatedly.

Missing Data Mechanisms

In a seminal paper, Rubin presented a framework for understanding missing data and their impact on statistical inference[3]. Here, I briefly describe the three categories of missing data formalized by Rubin.

MCAR: Data are **Missing Completely at Random** if the probability they are missing is unrelated to any variable under study. In the context of a randomized trial, unobserved values are on average distributed evenly across treatment arms and have little effect on estimates (although power may be reduced). That is, the observed values themselves comprise a simple random sample of the entire dataset. This is the strongest assumption one could make about the nature of the missing data because it presumes complete independence of the missing data mechanisms with other variables in the data. In practice, this unlikely. Under MCAR, simply excluding cases with incomplete observations will produced unbiased estimates.

MAR: Data are **Missing at Random** if the probability they are missing is related to another variable that has been completely observed. This is often a reasonable assumption to make in trials. For example, subjects may be more likely to drop out if their baseline measurements are worse than average. Under most conditions, data MAR are considered ignorable, that is, conditioned upon the observed data, the missing data mechanism can be ignored in lieu of modeling the full distribution of data and the probability of missingness. MCAR is a special case of MAR.

MNAR: Data are **Missing Not at Random** if the probability they are missing is related to the missing values themselves. In this case, even conditioning upon all the observed data, there are systematic differences between the distributions of observed and unobserved cases. Models that assume a MNAR mechanism require strong assumptions about the distribution of missing values that cannot be verified (as these values are unobserved). Hence, such models are often reserved for sensitivity analysis that aim to explore the sensitivity of conclusions to changing assumptions about the missing data mechanism.

Missing Data Patterns

In a data set that consists of multiple variables, as all randomized trials do, if missing data are present, then the pattern of missing data can play an important role in what assumptions are made about the missing mechanism. In addition, some patterns are more conducive to some statistical procedures than others. Restricting our attention to outcome variables in the context of a clinical trial, and where the outcomes variables are ordered over time (i.e. baseline and at least one other post-treatment measurement), a missing data pattern is called monotone when, for all cases, if a value is missing at one time, values are missing at all subsequent times. Conversely, a missing data pattern is intermittent, if a subject may be measured after a missed observation. Monotone missing data are often associated with drop out or withdrawal from a trial; although, this pattern may be evident in studies of terminal patients who die, or in trials with subjects who need rescue medication mid-trial and measurements are discontinued.

1.1.2 Non-inferiority Trials

There are some important differences between placebo-controlled or superiority studies and non-inferiority trials. First, one may view the designs in terms of their hypothesis statements being switched, so that the null condition of a superiority trial (no difference in effects between arms) is what is hoped to be shown in the non-inferiority trial. For instance, if μ_t is the mean effect in the new treatment arm and μ_c is the mean effect in control arm (placebo or active control), the hypothesis statements for a test of superiority are (assuming a larger effect is better):

$$H_0 : \mu_c - \mu_t = 0 \text{ versus } H_a : \mu_c - \mu_t < 0.$$

In comparison, a non-inferiority test has hypotheses:

$$H_0 : \mu_c - \mu_t > \delta \text{ versus } H_a : \mu_c - \mu_t \leq \delta.$$

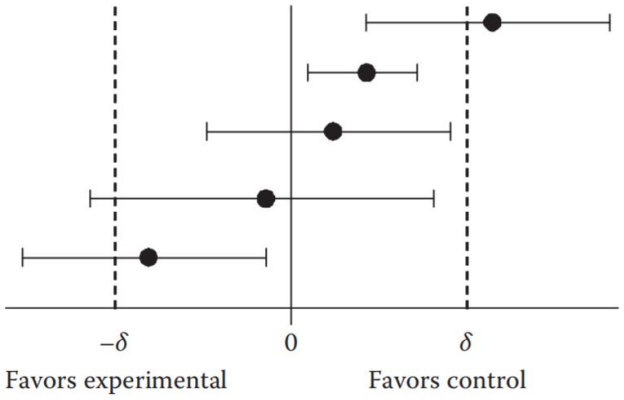

Figure 1.1: Point estimates and confidence intervals for control less experimental treatment in non-inferiority trials.[1]

The alternative in this case describes a mean experimental treatment effect that is at least as effective as the control less a little bit, $\delta\,(>0)$, the non-inferiority margin. Typically, a conclusion of non-inferiority is reached when the upper limit of a 95% confidence interval for $\mu_c - \mu_t$ is less than δ , which controls type I error at 2.5%. Figure 1.1 shows a selection of possible estimates and their confidence limits for non-inferiority trials and their conclusions.

Although, the idea of switched hypotheses is a useful heuristic, it is not precisely accurate because the non-inferiority alternative hypothesis includes the possibility that the new treatment is superior. In the superiority setting, the error control lies with the probability of rejecting a hypothesis of equivalence when the treatments are truly equivalent. Conversely, the error control in non-inferiority trials is on concluding treatments are equal when in fact they are not[4].

Another distinguishing feature of non-inferiority trials that differentiates them from superiority studies is the absence of a placebo arm. The consequence of having no placebo control means that a conclusion of non-inferiority does not automatically imply the treatment is effective. Given that showing efficacy of the new treatment is the ultimate, if not the direct, goal of the trial, the active control must perform as expected, but this cannot be directly measured[2]. Design and trial conduct are therefore critical to non-inferiority trials in reducing sources of error that could diminish trial validity[5]. Comparison of the results with historic studies or other means of validation could indeed be required to show compe-

tence of the study[4]. The International Conference on Harmonisation Statistical Principles for Clinical Trials E9 report also cites the lack of internal validity in non-inferiority trials and suggests the active control be chosen with care[6]. It should have been widely studied so that there is good evidence for its effectiveness (in placebo trials) and good estimates on its variability.

Non-inferiority Margin

One of the principle challenges in the design of non-inferiority trials is selection of the margin. Usually a margin is chosen to allow the experimental treatment to be slightly less effective than the active control. Furthermore, it is imperative that the margin be less than the effect of the active control as compared with placebo in historic studies. Indeed, often the margin is expressed as a percent effectiveness of the active control[2].

Unfortunately there is no accepted standard for selecting the margin and it cannot be validated or objectively determined[7]. At a minimum, the margin should be pre-specified in the trial protocol and justified on clinical grounds to accurately portray and honestly assess the new treatments performance. A systematic review of non-inferiority trials showed that methods for selecting the non-inferiority margin vary substantially, and their justification is not always reported[8].

The selection of the margin is critical because of its dependence on the effect of the active control. The effect of the active control needs to be maintained in the non-inferiority trial as it was in preceding studies for conclusions to be valid. The constancy assumption, which holds that conditions are similar to historic trials and the active controls treatment effect is preserved, if violated can result in a design with margin inappropriately large, where a non-inferiority conclusion does not imply effectiveness[5].

Assay Sensitivity and Constancy

Assay sensitivity is an experiments ability to detect meaningful differences between treatments. In an RCT, the property ensures that the trial design can adequately distinguish between effective and ineffective treatments. In non-inferiority RCTs, a trial with assay sensitivity could measure an effect of size M in the active control compared to a placebo

were a placebo arm included in the trial. Missing data could weaken the assumption of assay sensitivity[5]. For instance, patient dropout that is associated with positive outcomes, perhaps due to a perception of being cured, may attenuate the active control effect and, were a placebo arm included, not be shown effective. In this case, a conclusion of non-inferiority would not imply that the experimental treatment is superior to placebo.

The constancy assumption in non-inferiority trials supports assay sensitivity by assuming trial conditions are sufficiently close to those of historic trials in all ways that might influence the effect of the active control. Among the factors of importance are the inclusion/exclusion criteria, selection of analysis sets, and the analytic approaches used, all of which are intertwined with issues of missing data.

Analysis Sets/Estimands

The choice of analysis set needs to be considered carefully for non-inferiority trials given their anti-conservative nature. In the past, general guidelines for RCTs have advocated for intention-to-treat (ITT) analysis which ignore compliance information and analyze subjects as randomized. Recently, however, literature has focused on more thoughtful, pre-specified definition of estimands, which include a description of the population of interest, the outcome measures, handling of missing data and post-randomization events, and the type of analysis[9,10]. Nevertheless, there are concerns that estimands which target an ITT population for estimates of treatment effectiveness (as in the effectiveness of a treatment policy), can be anti-conservative. Garrett[4] suggested that, while the appropriate population for analyses is not always clear, the same population should be analyzed if switching hypotheses within a trial, for example, from non-inferiority to superiority.

Although selection of analysis sets is relevant whenever trial data have missing values or other protocol deviations, non-inferiority trials may indeed have some advantages over placebo-based RCTs when it comes to compliance. Trials with placebo arms may have greater challenges in preventing patient dropout or treatment switching because subjects know that they may have been allocated to placebo.

Bio-creep

Bio-creep is a phenomenon that occurs when, over time, a series of comparative trials gradually overstates the effectiveness of a treatment. Theoretically, in a series of non-inferiority trials where the active control was shown to be non-inferior to a different control in the previous non-inferiority trial, eventually, an ineffective drug may be deemed effective. Simulations have shown this is unlikely except with violations of the constancy assumption[11]; missing data could play a role in undermining constancy[12].

Other Issues

Some have criticized non-inferiority trials on ethical grounds by raising issues related to the value of conducting trials and subjecting patients to risk and inconvenience for treatments that are not expected to be superior. They argue that any external benefits of the new treatment may not be worth the additional risk for patients[7], and that such trials may overlook patients needs in favor of commercial interests[13].

Additionally, given that flaws in the design, conduct or analysis of non-inferiority trials will tend to favor the alternative conclusion, there is a greater risk of type I error. Hence, violations of entry criteria, treatment non-compliance, and missing data from withdrawals or drop out all complicate analysis and must be properly accounted for. This is significant because the cost of a type I error in non-inferiority trials could be higher than in superiority studies. There are two reasons for this. First, if an inferior treatment is taken when an effective one is available (i.e. the standard of care used as the active control), the opportunity cost is greater than it would have been if there were no effective treatment. Secondly, with no placebo we can infer the trialists could not deny treatment; hence, an inferior treatment would likely increase risk of mortality or disease progression compared with the active control.

1.1.3 Missing Data in Non-inferiority Trials

As previously alluded to, in the presence of missing data, the selection of analysis set has critical consequences for the conclusion of trials. If subject dropout is considered non-

compliant and imputed as such, estimands that target the population of all randomized participants can be biased towards the alternative. Similarly, non-compliance or treatment-crossover contamination can inflate type I error by estimating smaller differences between the effects in treatment arms. Alternatively, if an estimand targets the population of compliers with an interest in estimating biological efficacy, sometimes called a de jure estimand [10], the trialist may exclude non-compliers from the analysis. However, missing data further complicate the task: if subjects with missing data are included, they must be assumed to have complied and their missing values imputed as such. Alternatively, if they are excluded, as in a complete case analysis, the loss in randomization may introduce selection bias. In either case, the de jure estimand may not be generalizable.

If data are missing from non-inferiority data, how the missing mechanism affects the results will depend on how the data are handled and on whether there is interaction between the treatment arm and the missing mechanism. As in superiority designs, data MCAR can likely be excluded without risk of biased estimates.

Single imputation approaches to missing values, regardless of the underlying mechanism, are likely to introduce bias. Non-inferiority trials may be particularly susceptible to this bias unless the single imputation approach is conditional upon treatment arm. Otherwise, the imputed values are drawn from a single distribution which is consistent with the alternative hypothesis of non-inferiority. Indeed, any approach to missing data handling in non-inferiority trials that fails to account for a missing mechanism that varies between treatment arms is at risk of inflating type I error rates.

Sensitivity analyses examine whether conclusions change as a function of different assumptions about the missing data. In superiority trials, the reasonable range of assumptions may be different than those in non-inferiority trials. The U.S. Food and Drug Administration and others have proposed imputation models that assume the null hypothesis of inferiority, that is, a MNAR mechanism with different distributions for the observed and missing observations that vary by treatment arm [2,5]. Other approaches that may not be appropriate in superiority studies might introduce additional bias to the imputed data to restore some conservatism.

1.2 Approach

First, I performed a systematic review to study the prevalence of missing data and how they were being handled in the published literature of non-inferiority trials. This work is discussed in Chapter 3 and was published in the Journal of Pharmaceutical Statistics[14].

Secondly, I conducted a simulation study to characterize the effects of missing data in non-inferiority trials. I simulated data from a longitudinal non-inferiority clinical trial with missing values and treatment non-compliance under both the null and alternative hypothesis and assessed the performance of various statistical approaches for handling the missing values. The approaches were compared using measurements of error rates, power and bias, and I identified the characteristics of the data and missing mechanisms that were most likely to lead to biased results.

Finally, I developed a new multiple imputation model for sensitivity analysis in non-inferiority trials. Using information from the simulation study, I proposed a conservative imputation model that was easily implemented in existing software. I showed how the approach could be used as a sensitivity analysis investigating the impact of various plausible scenarios on the robustness of trial conclusions. The approach was demonstrated both within a tipping-point analysis framework and with using historic data on the active control arm to augment the posterior distribution of missing values with an informative prior.

1.3 Contributions

This research has advanced understanding of how missing data are handled, and on the issues affecting and solutions addressing inflated type I error in non-inferiority analysis. While regulatory agencies recommend statistically based methods, there has been a paucity of research on how these and other commonly used methods perform in non-inferiority trials and whether they lead to robust conclusions. Current recommendations by academic panels and regulatory agencies do not adequately address the statistical challenges of missing data unique to non-inferiority trials. For example, the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have recommended reporting both ITT and per-protocol analysis; however, this research supports the suggestion of some

authors that both analyses could be flawed in the presence of missing data. Reporting of both analyses should not increase confidence in conclusions.

Analysts have warned that non-inferiority trials with missing data may be biased towards concluding non-inferiority. This work improves understanding of how the analyses are most vulnerable to biased conclusions and makes specific recommendations for mitigating the risk of inflated type I error rates. Additionally, I have presented a new tool for examining the conclusions of non-inferiority trials that is easy to interpret and implement in existing software. The negative consequences of accepting an ineffective treatment cannot be overstated. At minimum, drug developers may incur unnecessary costs, and at worst, patients are put at risk. Ultimately, the contributions from this dissertation will improve reproducibility and minimize the risk of false conclusions from non-inferiority trials with missing data.

Chapter 2

Literature Review

The literature on missing data in randomized clinical trials (RCTs) is extensive[15–17]. Methodologists have delved into their effects and developed new procedures for their handling. Theoretical results, in addition to simulation studies have shown that missing outcome data can reduce statistical power, and unless the data are MCAR, can often lead to biased estimates of effect size[18–20]. Due to their growth in popularity, non-inferiority trials have been the subject of new research on missing data and protocol deviations, particularly with respect to difficulties in control of type I error. Here we give some example of relevant studies addressing the challenges.

Wiens and Rosenkranz[5] conducted simulations to investigate type I error, bias and power using several approaches to missing data handling in non-inferiority trials. Their approaches included analysis with linear mixed models and a less principled imputation method, last-observation-carried-forward (LOCF), for longitudinal continuous outcomes in non-inferiority trials. They concluded that linear mixed model performs well when data are MAR and sometimes MNAR. That is, there is support for the conclusion that missing data can be handled just as well with a mixed model in non-inferiority trials as they are in superiority studies. They stated a need for further research on multiple imputation models and on sensitivity analyses when the MAR mechanism is not believed to adequately explain the missingness.

A heavily cited study from 2006 investigated the selection of analysis sets in non-inferiority trials[21]. Sanchez and Chen simulated continuous longitudinal outcomes with

missing data and several types of protocol deviations including non-compliance. For imputations in their primary ITT analysis, they used LOCF, and the per-protocol analysis excluded cases with protocol deviations including missing observations as in a complete-case analysis. Results from the ITT and per-protocol analyses were obtained using t-tests and compared for bias and type I error. In addition, the authors compared results from a linear mixed model on MAR data and found it to be more efficient that the per-protocol analysis of complete cases. An important finding was they identified scenario in which the ITT was conservative and the per-protocol was anti-conservative, contradicting the conventional understanding of these analysis sets in non-inferiority settings. The authors suggested that a hybrid per-protocol/ITT analysis which excludes non-compliant patients and makes more reasonable assumptions about missing data will improve model estimates. Yoo [22] simulated non-inferiority trial data and performed an analysis of covariance (ANCOVA) with LOCF imputation, linear mixed models and a pattern-mixture model approach. The author concluded that there is no single best statistical practice for avoiding bias and controlling type I error rates under a broad range of realistic MAR and MNAR scenarios. Lipkovich and Wiens [23] investigated binary outcomes with multiple imputation for missing longitudinal covariate data and compared various imputation models. They found that, as in superiority studies, multiple imputation performs well when important variables associated with the missing outcomes and probability of missingness are part of the imputation model, including the treatment arm assignment.

Practical guidelines from regulatory agencies have acknowledged the design and analysis challenges to non-inferiority trials and have urged researchers to take measures to limit or minimize the harms of missing data [15]. In guidance documents released by The U.S. Department of Health and Human Services, Food and Drug Administration, missing data are mostly addressed in the context of selecting the analysis set [2]. The FDA cites several characteristics of study quality that can lead to a false conclusion of non-inferiority including poor compliance, use of concomitant treatments, inadequate measurement instruments or techniques, high attrition and poor follow-up. Many of these issues can be viewed through the lens of missing data problems. The document also emphasizes the potential for anti-conservatism of the ITT analysis which is viewed as conservative in superiority studies. One

suggestion for improving control of type I error is to use an imputation model that assumes the null hypothesis. This is congruent with suggestions by Wiens and Rosenkranz [5], who proposed imputing treatment arms using different values, but perhaps not as extreme as those under the null.

The FDA also addressed issues of constancy and assay sensitivity in non-inferiority trials, recognizing the importance of knowing whether the active control has had its expected effect. The guidelines do not relate these problems directly to missing data. The International Conference on Harmonisation [6] discusses the role of estimands and sensitivity analysis in clinical trial. With regards to non-inferiority designs, the guidelines caution the trialist that selection of primary estimands (which include the choice of the analysis set) must be considered carefully due to the anti-conservative nature of non-inferiority and equivalence trials. Furthermore, it regards the ITT analysis (which they call the full analysis) as one that is inherently anti-conservative but stops short of calling for a per-protocol analysis. The document emphasized construction of estimands which target the specific reasons treatments could be made to appear more similar that are not related to real differences in efficacy between treatments, that is, any event (like receiving rescue medication) that could lead to an attenuation of differences between arms. The European Agency for the Evaluation of Medicinal Products advises that the per-protocol and ITT analysis be considered equally and reach the same conclusions for a robust analysis [24].

Therefore, efforts to improve analysis and reporting have included conducting both the ITT and per-protocol analysis and examining them for consistency [25]. However, simulations have shown these analyses can be biased toward the same conclusion. Furthermore, some argue that requiring trialist to report two estimands can be burdensome on decision makers [12]. A review of 20 antibiotic non-inferiority trials reporting both ITT and per-protocol estimands could not determine whether estimates of effect size were associated with choice of analysis set [26]. The authors speculated that both analyses may often underestimate the true differences between the active control and experimental treatment. Some authors have cautioned against using per-protocol estimands as they have flavor of cherry-picked data. In a superiority trial setting, it is obvious that removing cases for non-compliance can overstate effect size, particular if the reason for non-compliance is related to efficacy [12]. One could

make a stronger argument for analyzing only the adherent subjects when the objective is to show non-inferiority, especially if there are equal rates of non-adherence between arms.

Gamalo et al.[27] proposed a Bayesian approach to defining the non-inferiority margin using meta-analysis of historical data and Dirichlet process priors. They also presented a Bayesian decision rule for analysis and demonstrated sample size calculations using this rule. The authors noted missing data as a potential source of bias and suggested comparing results from frequentist and Bayesian analyses to better understand the observed treatment effects. Viele et al.[28] discussed various approaches to utilizing historic data, particularly on the active control, to improve precision and reduce bias of estimates. Their findings suggested that the improvements to estimates from historic borrowing would often depend upon the similarity of the current study to historic studies.

Some authors have used an enhanced tipping point analysis for two-armed clinical trials with a graphical display of conclusions. Liublinska and Rubin[29] demonstrated a tipping point analysis of all possible response rates for the missing binary outcomes in a two-arm trial. For every possible proportion of response among the subjects with missing data for each arm, p-values or estimates could be displayed in a two-dimensional grid. This approach was touted as a practical approach to sensitivity analysis with respect to missing data assumptions that is easily interpretable by clinicians and helpful for evaluating plausible missing mechanisms. Kim et al.[30] applied the enhanced tipping points to a sensitivity analysis for non-adherence in non-inferiority trials. Their approach was to address the inherent difficulty of selecting the analysis set by conducting a sensitivity analysis with tipping points that shows results for counterfactual estimands for all possible assumptions about patient compliance. In Chapter 5, I adopt the enhanced tipping point analysis to display results from a sensitivity analysis for missing continuous outcomes.

Chapter 3

Systematic Review of Missing Data Handling in Non-Inferiority and Equivalence Trials

3.1 Motivation

Systematic reviews of clinical trials have shown that missing data are prevalent and often not handled in a statistically based manner[16,17,31]. Although there has been an acknowledgement in the literature that protocol deviations can be particularly deleterious, few reviews of trial reporting and conduct have focused on missing data. Although some reviews of non-inferiority trials had assessed quality of reporting, especially with respect to justification of the non-inferiority margin, none had focused on the important issues of missing data nor assessed statistical practices. To fill the gap, we conducted a review of non-inferiority trials to estimate the proportion of missing data in the design and to discover the commonly used methods for handling missing outcomes. Furthermore, with the recent issuance of new guidelines from the FDA and the National Research Council Panel on Missing Data, a review of current practices was timely and important for comparison with older reviews to see whether the guidelines had any impact on statistical practice.

3.2 Methods

We reviewed a sample of non-inferiority trials for the extent, handling, and use of sensitivity analysis for missing data. We included equivalence trials, which aim to show two treatments are therapeutically similar with a pre-defined equivalence margin. Equivalence trials are less common, but they share the potential for anti-conservatism with non-inferiority designs. We also investigated the choice of analysis set, whether per-protocol, ITT or both. Our secondary aims were to explore other trial characteristics such as the quality of the design and reporting, and we compared our findings between journals with high and low impact factors.

The review consisted of randomly selected journal articles from a PubMed database search using the keywords "non-inferiority", "active control", "equivalence", "equivalency", and "statistically equivalent". Only trials published in English between May 1, 2015 and April 30, 2016 were considered. We excluded trials with primary survival outcomes because these raise different issues due to the censoring of missing data[16]. We also excluded pilot and observational studies, secondary reports, and bioequivalence studies. We oversampled from five top medical journals, Annals of Internal Medicine, BMJ, JAMA, The Lancet, and The New England Journal of Medicine so that we could compare the results of our review with previous reviews.

We collected data on the number of proportion of patients with missing data in the primary outcome, using the number of missing values at the final timepoint to estimate proportion in the case of repeated measures. Most of the missing data information, such as numbers and cause of missingness if reported, was extracted from the CONSORT flow diagram[32]. Otherwise, the data were obtained by a thorough reading of the results.

We recorded the primary analysis method as well as any missing data handling methods that were explicitly cited: complete case analysis, single imputation methods, mixed models or multiple imputation. If missing data handling was not discussed, we assumed a complete case analysis was conducted (unless a mixed model was used for the primary analysis). Also, we noted whether sensitivity analysis was done and information about the analysis sets used, whether ITT, per-protocol or both.

Other data we collected on the trial characteristics included details on type of trial (drug, device or other), type of outcome, sample size calculation, justification for the non-inferiority margin, study conclusions, journal, impact factor and source of funding.

3.3 Main Findings

We reviewed 109 trials and found that 93% reported some missing outcomes. Almost half reported more than 10% missing values in the final outcome. The mean proportion of missing values was 10.6%. Half of the trails reviewed excluded cases with missing data (complete case analysis) and 28% used a single imputation approach such as the "last observation carried forward", mean replacement, or worst-case imputation. Only 12% of trials reviewed used a model-based method such as a mixed model or multiple imputation for handling missing data.

The proportion of missing data was consistent with previous reviews of missing data in other types of RCTs. Sensitivity analysis examining the impact of assumptions about the mechanism were reported in 11 articles, lower than reported on other systematic reviews.

A third of articles reporting results for ITT and per-protocol analyses both in accordance with regulatory guidelines. Often, no adjustment was made to the sample size in anticipation of subject dropout, and many did not justify the choice of margin. Additionally, very few authors discussed issues related to the dependence of conclusions on the effect of the active control. Articles published in high impact journals we associated with higher quality reporting: they were more likely to conduct sensitivity analysis.

The systematic review was published in the Journal of Pharmaceutical Statistics and is reproduced in appendix A.

Chapter 4

A Conservative Approach for Analysis of Non-Inferiority Trials with Missing Data and Subject Non-Compliance

Abstract

Non-inferiority clinical trials aim to show an experimental treatment is therapeutically no worse than standard of care, particularly if the new treatment is preferred for reasons such as cost, convenience, safety, and so on. Non-inferiority trials are by nature less conservative than superiority studies: protocol violations may increase bias toward the alternative hypothesis of non-inferiority. Our objective was to compare multiple imputation, a linear mixed model, and other methods for analyzing a longitudinal trial with missing data in intention-to-treat and per-protocol populations. We simulated trials with missing data and non-compliance due to treatment inefficacy under varying trial conditions (e.g. trajectory of treatment effects, correlation between repeated measures, and missing data mechanism), assessing each approach by estimating bias, type 1 error and power. We found that multiple imputation using auxiliary data on non-compliance in the imputation model performed best. A hybrid intention-to-treat/per-protocol multiple imputation approach with

a control-based, missing not at random imputation model produced low type 1 error, was unbiased and maintained reasonable power to detect non-inferiority. We conclude that the anti-conservatism of non-inferiority trial estimands conforming with the intention-to-treat principle may be offset by imputation models that include variables on intercurrent events.

4.1 Introduction

Non-inferiority clinical trials test the hypothesis that a new treatment is not unacceptably worse than the existing standard of care by some pre-defined margin. If a new treatment has potential advantages related to cost, safety, toxicity, convenience, etc., it may be reasonable to accept a new treatment even if it slightly underperforms the standard of care in efficacy [2].

Among the important issues in the design of non-inferiority trials is the choice of analysis set, especially when a trial is anticipated to have subject dropout or protocol violations [33]. Even when researchers continue collecting data on those who break treatment protocol, questions remain as to how analysis of off-treatment subjects should proceed. One possibility is to select the per-protocol analysis set and exclude all those who have discontinued or otherwise broken protocol, which may include removing subjects with missing data. While the types of protocol violations are numerous, some post-randomization events may be related to outcome (e.g., treatment switching, receipt of rescue medication, and non-compliance due to treatment inefficacy), in which case the per-protocol estimator is subject to selection bias. The intention-to-treat principle is conventionally employed to remedy this problem by analyzing subjects as randomized regardless of protocol adherence. Estimands based on the intention-to-treat principle preserve the balance of covariates between arms and allow for broader interpretability. While generally conservative in superiority studies, in non-inferiority trials, intention-to-treat analysis may bias results towards the alternative conclusion of non-inferiority [12]. For example, inclusion of non-compliant subjects may attenuate estimates of the difference between arms [21].

Rubin [3] defined three primary missing data mechanisms which we frame in terms of missing outcome data. Data missing completely at random (MCAR) assumes that the probability that outcome data are missing is independent from all observed and unobserved

data. Truly MCAR data are often benign and at worst lead to a reduction in power. When outcome data are missing at random (MAR), the probability of missingness is based on some previously observed data. Statistically principled methods for handling MAR data include likelihood-based methods (provided models are properly specified), multiple imputation and inverse probability weighted generalized estimating equations. Finally, outcome data are missing not at random (MNAR) when the probability of being missing depends on the unobserved data, even after conditioning on observed data. There is no way to distinguish MAR from MNAR data, therefore sensitivity analyses should be used to explore how departures from assumptions about the missing data affect the trial estimates and conclusions.

Missing data present challenges to the analysis of all clinical trials. However, non-inferiority trials are particularly vulnerable to inflated type I error rates due to missingness. Estimands that target the population of all randomized subjects, in adherence with the intention-to-treat principle, often require that missing values be imputed. Hence, if a trial has missing data, intention-to-treat analysis may inflate the type 1 error rate if unobserved values are not imputed in a thoughtful manner[5]. For example, the commonly used last-observation-carried-forward approach has the reputation of conservatism in superiority studies; however, last-observation-carried-forward can easily lead to anti-conservative estimates when testing for non-inferiority[21].

There are two practical consequences of false positives in non-inferiority trials which may be more severe than in superiority trials. First, if an inferior treatment is taken when an effective treatment is available, the opportunity cost is greater than it would have been if there were no effective treatment (and this may be so in the case of a placebo-controlled superiority study.) Second, the decision to omit a placebo arm from the trial often implies that it would be unethical to forgo standard of care. Hence, using an inferior treatment would likely lead to greater risk of mortality or morbidity than were the active control used as treatment instead.

The statistical literature has exposed some of the difficulties related to missing data in non-inferiority trials. Sanchez and Chen[21] investigated the choice of analysis set. They found that the conservatism or anti-conservatism of analysis was highly dependent on the

types of missing data and protocol violations. Furthermore, they noted non-compliance often leads to anti-conservative conclusions. While their study provides meaningful insight into longitudinal non-inferiority trials with continuous outcomes, they did not investigate multiple imputation or data MNAR. Wiens and Rosenkranz [5] concluded that the likelihood-based mixed model for repeated measures, described in detail below, was effective in handling data MAR, even in some cases when data were MNAR, supporting the view that likelihood based methods are appropriate tools for the non-inferiority setting. However, the authors noted the lack of research on multiple imputation as a potential sensitivity analysis. Lipkovich and Wiens [23] showed that multiple imputation is effective at imputing missing longitudinal covariate data with a binary primary outcome in the non-inferiority setting, supplying further evidence that multiple imputation is most useful when data are MAR and important variables are included in the imputation model.

Our research investigates missing data and the specific intercurrent/post-randomization event of non-compliance due to treatment inefficacy, and how these may bias conclusions toward the alternative hypothesis. We aimed to find a conservative approach that balances the benefits of preserving randomization with the desire to obtain estimates of the true biological difference in treatment effects. We evaluated several methods of missing data handling compatible with various intention-to-treat and per-protocol estimands including multiple imputation, a linear mixed model, and the commonly used last-observation-carried-forward imputation and complete cases analysis [14]. Through simulations, we compared type 1 error rates, bias and power for each approach in a longitudinal non-inferiority trial with continuous outcomes. We also aimed to identify the main characteristics of the missing data mechanism that lead to biased estimates, and to explore how other features of the data, such as the trajectory of treatment effects and the correlation between repeated measures, bias trial estimates across a range of missing data proportions. We demonstrate our methods using an example from a dose-finding study of an estrogen suppression drug in post-menopausal women with increased risk of breast cancer.

4.2 Methods

4.2.1 Overview

Data were simulated to represent outcomes from a balanced two-arm, non-inferiority trial with measurements collected at baseline and at three subsequent time points. A proportion of the simulated subjects was selected for non-compliance and their measurements at the final time point were attenuated. Outcome data were removed under various missing data mechanisms. Throughout, we assumed that compliance status on every subject was observed, even for those with missing outcome data.

We simulated multivariate-normal outcome data, and without loss of generality, assumed that larger treatment effects were desirable. The parameter of interest, Δ, was the difference between treatment effects, μ_{AC} and μ_T, defined as the mean change from baseline at the final time point in the active-control (AC) and experimental treatment (T) arms respectively. For simulations under the null hypothesis, we set $\mu_{AC} = 2$ and $\mu_T = 1$, and under the alternative, $\mu_{AC} = 2$ and $\mu_T = 2$. The trial hypotheses were: $H_0 : \Delta = \mu_{AC} - \mu_T > \delta$ versus $H_a : \Delta : \mu_{AC} - \mu_T \leq \delta$, where the non-inferiority margin, δ, was chosen to be 1. We varied the following factors: the outcome's trajectory over time, the within-subject correlation, the proportion of missing data, the missing data mechanism, and the direction of dropout (described below). In addition, data were generated under the assumptions of full compliance with trial protocols and with 80% of subjects compliant resulting in attenuated treatment effects in both arms. A summary of simulation factors is shown in the supplemental materials.

4.2.2 Data Generation

Outcome data were generated from a multivariate normal distribution with means determined by treatment arm. For each subject i, $(i = 1, , N)$, we simulated outcomes, y_{ij}, at each time j, $(j = 0, ..., 3)$, where the vector of outcomes, $\mathbf{Y_i} = (y_{i0}, y_{i1}, y_{i2}, y_{i3})$, was given by $\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta_i} + \boldsymbol{\epsilon_i}$ with $\boldsymbol{\epsilon_i} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. We assumed a heterogeneous autoregressive

variance-covariance structure for the repeated measures on subjects:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho \\ & & & \sigma_4^2 \end{pmatrix}$$

with $\sigma_1 = 1.5$,$\sigma_2 = 1.75$,$\sigma_3 = 2.25$, and $\sigma_4 = 2.5$. We simulated data using two values for the correlation parameter, ρ: 0.3 and 0.7. Under the null, the trajectories over time were modelled by two different non-linear patterns. The first, which we label trajectory 1, was characterized by early separation of the AC and T arms at the first post-baseline measurement so that $(\boldsymbol{\beta_i}^{EARLY})^T = (0,\ 0.6,\ 0.9,\ 1)$, for i in T, and $(\boldsymbol{\beta_i}^{EARLY})^T = (0,\ 0.9,\ 1.5,\ 2)$ for i in AC. In the second pattern, trajectory 2, trajectories separated late: $(\boldsymbol{\beta_i}^{LATE})^T = (0,\ 0.8,\ 1.5,\ 1)$, for i in T, and $(\boldsymbol{\beta_i}^{LATE})^T = (0,\ 0.8,\ 1.5,\ 2)$ for i in AC. Under the alternative hypothesis, the effects over time in AC and T were equal with means $(\boldsymbol{\beta_i}^{H_A})^T = (0, 0.9, 1.5, 2)$ for every $i = 1, ..., N$. Sample size was fixed at 210 subjects per arm to achieve 90% power to show non-inferiority with a margin of 1, while controlling type I error at 5%[34].

Non-Compliance

The first stage of simulations assumed fully compliant subjects with no modification of effect due to protocol violations. However, we wished to assess the additional impact of non-compliance due to inefficacy. Our model selected subjects as "candidate" non-compliers if the measured effect at follow-up time 2 did not exceed the mean of the first two measurements (baseline and time 1) by at least 1 unit. The probability of non-compliance increased in proportion to the difference between time 2 and the mean of the previous two times. Let C_i be an indicator variable taking the value 1 if subject i does not comply and 0 otherwise. Then $P(C_i = 1|y_{i2} - (y_{i0} + y_{i1})/2 < 1) \propto ((y_{i0} + y_{i1})/2 - y_{i2})$. Those selected were modified to reflect a 50% reduction in treatment effect at time 3 (the final assessment). A tuning parameter was used to achieve an approximate 80% compliance rate.

Missing Data

After generating the complete data and attenuating outcomes for non-compliance, we simulated MAR and MNAR data by removing values at times 2 and 3. We defined the proportion, p, of missing data as the number of missing observations at the final time point divided by the total number of subjects. For the primary simulations, in which we analyzed 1000 replicates of every trial scenario, we chose $p = 0.2$ to be consistent with the proportions reported in real longitudinal clinical trials[14,17]. We also allowed p to vary between five and 50 percent in approximately 1% increments, simulating 100 replicates for each scenario.

To produce data MAR, for subject i at times $j = 2, 3$, y_{ij} was removed with probability proportional to the z-score of the previous observation. Given a missing indicator R_{ij} that is equal to 0 if y_{ij} is missing and 1 if observed, then $P(R_{ij}^{MAR} = 0) \propto (\bar{y}_{j-1} - y_{i,(j-1)})/s_{j-1}$. where $\bar{y}_{j-1}$and s_{j-1} were the sample mean and standard deviation of observations in both treatment arms at time $j-1$. Lower values of $y_{i,(j-1)}$ were more likely to cause missing y_{ij}. MNAR data were removed in a similar manner to MAR except that missingness probability depended on current time points, that is, $P(R_{ij}^{MNAR} = 0) \propto (\bar{y}_j - y_{ij})/s_j$. Subjects with missing data at any time point had all measurements at subsequent times set to missing to create a monotone missing pattern. Tuning parameters were used to obtain the correct proportion of missing data.

To account for various factors contributing to the cause of missing data, we also considered a scenario in which data were missing in opposite directions between treatment arms, that is, a treatment arm interaction with the missing mechanism. In this scenario, which we refer to as "opposite direction" missingness, smaller outcomes in the T arm were more likely to be missing whereas larger outcomes in the AC arm had a higher probability to be missing. This scenario in a real trial is plausible and might arise if the toxicity of an experimental drug tends to overshadow its effectiveness and cause poor responders to drop out while among those taking the active control, a high response is associated with dropout because subjects have the experience of feeling "cured". The missing mechanism with no interaction we will refer to as "same direction" missingness.

4.2.3 Analysis and Evaluation

We evaluated a total of six estimation procedures which differed by at least one element: their analysis set, how missing data were handled, and the primary analysis. On data simulated without the non-compliance factor, we compared only three of the six procedures:

(1) intention-to-treat with last-observation-carried-forward imputation and t-test

(2) intention-to-treat with mixed model for repeated measures[18]

(3) per-protocol, complete case analysis and t-test.

The last-observation-carried-forward approach for longitudinal clinical trials has been described previously[12]. The mixed model for repeated measures, which utilizes information about correlation of successive measurements, allows subjects with dropout to contribute to the estimation at the final time point. The model is robust to data MAR (when variables associated with missingness are in the model) and with a correctly specified model. Time is defined categorically to allow for non-linearity. In addition, the covariance matrix of the repeated measures is determined completely from the data (unstructured). The model was fit using restricted maximum likelihood estimation.

When we simulated data with some subjects non-compliant, six procedures were evaluated. We compared the aforementioned procedures plus an additional three:

(4) intention-to-treat with multiple imputation and mixed model

(5) per-protocol with mixed model

(6) an intention-to-treat/per-protocol hybrid with control-based multiple imputation.

All of the per-protocol analyses excluded every non-compliant subject. In the multiple imputation procedure (4), we used a regression method which assumes multivariate normality of imputed variables with observed outcomes, treatment arm and a non-compliance indicator variable in the imputation model. We generated ten imputations, analyzed each with the mixed model and pooled estimates according to Rubins rules[35].

The hybrid intention-to-treat/per-protocol analysis (6) considered data collected on non-compliant subjects akin to missing. Hence, data at the final time point for non-compliant

subjects were removed. Subsequently, within a multiple imputation framework, we imputed values (removed both via the missing data mechanism and for non-compliance) using control-based pattern imputation under the MNAR assumption[36,37]. We achieved this by restricting the observations from which the imputation model is derived to only subjects who are compliers. This procedure is consistent with a hypothetical estimand described in the International Conference on Harmonisation[6] E9 Report that targets the biological difference in effects in the population of subjects who adhere with protocol. SAS code for each of the models is shown in the supplemental materials.

Under the null hypothesis, type I error was estimated by the mean number of trial iterations in which the upper limit of the 90% confidence interval fell below the non-inferiority margin of 1, so that the expected type 1 error rate was 5%. Estimated percent bias, defined as (estimate-1)*100, was reported as well. Under the alternative hypothesis, we estimated power to detect $\Delta = 0$. Power was estimated by the proportion of replicates resulting in a conclusion of non-inferiority (where the lower limit of the 90% confidence interval was greater than $\delta = 1$).

4.3 Results

Table 4.1 shows the estimates of effect, percent bias, and type 1 error under the null hypothesis of inferiority when the true change from baseline difference between arms is 1. The data were simulated to have approximately 20% missing at the final time point, a monotone missing pattern, full compliance, and with moderate correlation between the repeated measures ($\rho = 0.7$). Overall, the mixed model for repeated measures outperformed other methods by controlling type 1 error and producing unbiased estimates when the missing mechanisms was MAR. However, under MNAR, type 1 eror was inflated, most severely with interaction between the missing mechanism and treatment arm. Last-observation-carried-forward performed poorly, with inflated type 1 error rates, and bias that included underestimation of the treatment effect by approximately 2 to 50%, as well as overestimation by 7%. The estimates produced from the last-observation-carried-forward approach were also most sensitive to the trajectory of treatment effects of all the methods: type 1 error rates under

trajectory 1 (early separation of effects) were often greater than those under trajectory 2 by over 10%. Complete case per-protocol analyses performed reasonably when the missingness direction was the same by arm but was biased with inflated type 1 error when the direction was opposite. Results for simulations with low correlation between repeated measures were similar.

Table 4.1: Full Compliance and 20% Missing Data. Estimates (percent bias) of difference in change from baseline and type I error under the null hypothesis of inferiority with 20% missing data at final time point. Estimates are difference in change from baseline between arms at final timepoint. Correlation between the repeated measures is $\rho = 0.7$. A : Last Observation Carried Forward. B : Mixed Model for Repeated Measures. C : Type 1 Error. D : Trajectory 1 indicates the early separation of effects by treatment arm; in trajectory 2, the effects separate late.

	Intention to Treat				**Per Protocol**	
	LOCF[A]		**MMRM[B]**		**Complete Cases**	
	Estimate (% bias)	T1E[C]	Estimate (% bias)	T1E	Estimate (% bias)	T1E
MAR - Same						
Trajectory 1[D]	0.80 (-19.7)	21.20	1.00 (0.2)	4.7	1.00 (-0.1)	5.9
Trajectory 2	0.92 (-7.6)	8.3	1.01 (0.7)	4.7	0.98 (-1.6)	5.2
MAR - Opposite						
Trajectory 1	0.98 (-1.5)	4.7	1.00 (-0.2)	5.6	0.75 (-25.4)	25.5
Trajectory 2	1.07 (7.1)	3.5	1.00 (0.3)	6.3	0.75 (-25.4)	25.9
MNAR - Same						
Trajectory 1	0.77 (-23.2)	27.9	0.96 (-3.5)	7.5	0.95 (-5.0)	8.1
Trajectory 2	0.88 (-11.5)	14.0	0.94 (-5.9)	8.2	0.92 (-7.7)	9.5
MNAR - Opposite						
Trajectory 1	0.49 (-51.4)	76.6	0.45 (-55.4)	72.6	0.30 (-70.3)	85.1
Trajectory 2	0.57 (-42.8)	61.8	0.44 (-55.8)	71.8	0.29 (-71.0)	85.9

With the introduction of non-compliant subjects, simulation results under all 6 methods are shown in Table 4.2. So we could assess bias from non-compliance separately from the bias induced by missingness, the first two rows of Table 4.2 show estimates and type 1 error for the complete data (including non-compliers), analyzed with the mixed model for repeated measures for all subjects (intention-to-treat), and for only compliant subjects (per-protocol). We also analyzed the complete data under the intention-to-treat/per-protocol hybrid where values at the final timepoint of non-compliant subjects were set to missing and imputed. These results show that, in the absence of missing data, the intention-to-treat analysis underestimated the true change from baseline difference in effects by approximately 10%

and type 1 error was inflated to 10%. Our intention-to-treat/per-protocol hybrid approach produced unbiased estimates and achieved type 1 error near 5%.

With 20% missing data, last-observation-carried-forward performed worst among methods, repeating the patterns seen in Table 4.1. The intention-to-treat mixed model and multiple imputation approaches were anti-conservative as a direct result of bias from non-compliance. In the per-protocol analysis, bias was more sensitive to the trajectory of effects. Indeed, when the mean trajectories separated early, the per-protocol analysis produced less biased estimates than under trajectory 2 with late separation. The hybrid analysis with MNAR multiple imputation produced unbiased estimates consistently when data were MAR, regardless of the trajectory of effects. Furthermore, type 1 error rates were lower than that of other methods. The results for low correlation between repeated measures were similar.

As seen in Tables 4.1 and 4.2, among all the missingness mechanisms, the MNAR pattern with treatment arm interaction, missingness in opposite directions, produced dramatically biased estimates. This effect is seen in Figure 4.1 which compares estimated percent bias as a function of percent missing data from selected analyses: intention-to-treat with mixed model, per-protocol complete case analysis, and hybrid intention-to-treat/per-protocol with multiple imputation. Under the data MNAR in opposite directions, percent bias was strongly associated with the proportion of missing data, and the treatment effect was underestimated with increasing percent missing. Other missingness patterns produced less bias even with high levels of missingness. Except for the MNAR-opposite pattern, the hybrid approach performed well when data were MAR at even high proportions of missing data.

A summary of type 1 error rates is shown in Figure 4.2. The per-protocol, complete case analysis was inconsistent in controlling type 1 error rates. When the mixed model for repeated measures was used on the per-protocol analysis set, type 1 error was better controlled and more conservative than the mixed model applied to the intention-to-treat analysis set. The MNAR-opposite direction pattern produced very large type 1 error rates. Among the six approaches, the hybrid MNAR multiple imputation approach was most conservative with the lowest overall type 1 error rates across a range of missing data proportions.

Table 4.2: 80% Compliance and 20% Missing Data. Estimates (percent bias) and type I error under the null hypothesis of inferiority with 20% missing data at final time point. Estimates are difference in change from baseline between arms at final timepoint. Correlation between the repeated measure is $\rho = 0.7$. A : Last Observation Carried Forward. B : Mixed Model for Repeated Measures. C : Multiple Imputation. D : Multiple Imputation with MNAR Control-Pattern. E : Type 1 Error. F : Trajectory 1 indicates the early separation of effects by treatment arm; in trajectory 2, the effects separate late.

	Intention to Treat					
	LOCF[A]		**MMRM[B]**		**MI[C]**	
	Estimate (% bias)	T1E[E]	Estimate (% bias)	T1E	Estimate (% bias)	T1E
No Missing						
Trajectory 1[F]	-	-	0.92 (-7.8)	10.0	-	-
Trajectory 2	-	-	0.92 (-8.2)	9.8	-	-
MAR - Same						
Trajectory 1	0.75 (-24.5)	28.9	0.94 (-6.2)	8.1	0.93 (-7.4)	8.8
Trajectory 2	0.88 (-11.7)	12.3	0.96 (-4.3)	7.6	0.94 (-6.0)	8.3
MAR - Opp.						
Trajectory 1	0.96 (-4.2)	7.5	0.93 (-7.0)	9.0	0.93 (-7.2)	9.5
Trajectory 2	1.06 (5.9)	2.8	0.94 (-5.6)	8.4	0.92 (-7.7)	8.9
MNAR - Same						
Trajectory 1	0.71 (-29.1)	38.3	0.89 (-11.1)	11.9	0.89 (-10.6)	10.1
Trajectory 2	0.85 (-14.5)	16.6	0.90 (-9.7)	10.4	0.90 (-9.7)	10.6
MNAR - Opp.						
Trajectory 1	0.47 (-52.7)	79.4	0.40 (-59.7)	78.9	0.46 (-53.7)	70.3
Trajectory 2	0.57 (-43.4)	65.0	0.41 (-59.4)	79.7	0.50 (-50.30)	66.0

	Per Protocol				**Hybrid**	
	Complete Cases		**MMRM**		**MI - MNAR Control[D]**	
	Estimate (% bias)	T1E	Estimate (% bias)	T1E	Estimate (% bias)	T1E
No Missing						
Trajectory 1[F]	-	-	1.01 (0.6)	5.0	1.00 (-0.1)	4.8
Trajectory 2	-	-	0.89 (-10.9)	11.8	0.99 (-1.1)	4.5
MAR - Same						
Trajectory 1	1.01 (0.9)	5.6	1.01 (0.8)	4.6	0.99 (-0.8)	5.4
Trajectory 2	0.90 (-10.1)	9.4	0.91 (-9.4)	9.2	0.98 (-1.7)	5.4
MAR - Opp.						
Trajectory 1	0.81 (-18.8)	17.5	1.00 (-0.1)	5.6	1.00 (-0.16)	5.6
Trajectory 2	0.74 (-25.8)	23.9	0.92 (-8.1)	9.9	0.98 (-1.6)	6.7
MNAR - Same						
Trajectory 1	0.93 (-7.2)	9.5	0.94 (-5.5)	8.9	0.95 (-5.1)	6.7
Trajectory 2	0.85 (-15.2)	13.6	0.86 (-14.4)	12.6	0.91 (-8.9)	10.5
MNAR - Opp.						
Trajectory 1	0.41 (-59.3)	70.5	0.50 (-50.0)	61.4	0.44 (-56.1)	66.7
Trajectory 2	0.31 (-69.2)	83.1	0.39 (-60.6)	74.9	0.44 (-55.5)	64.6

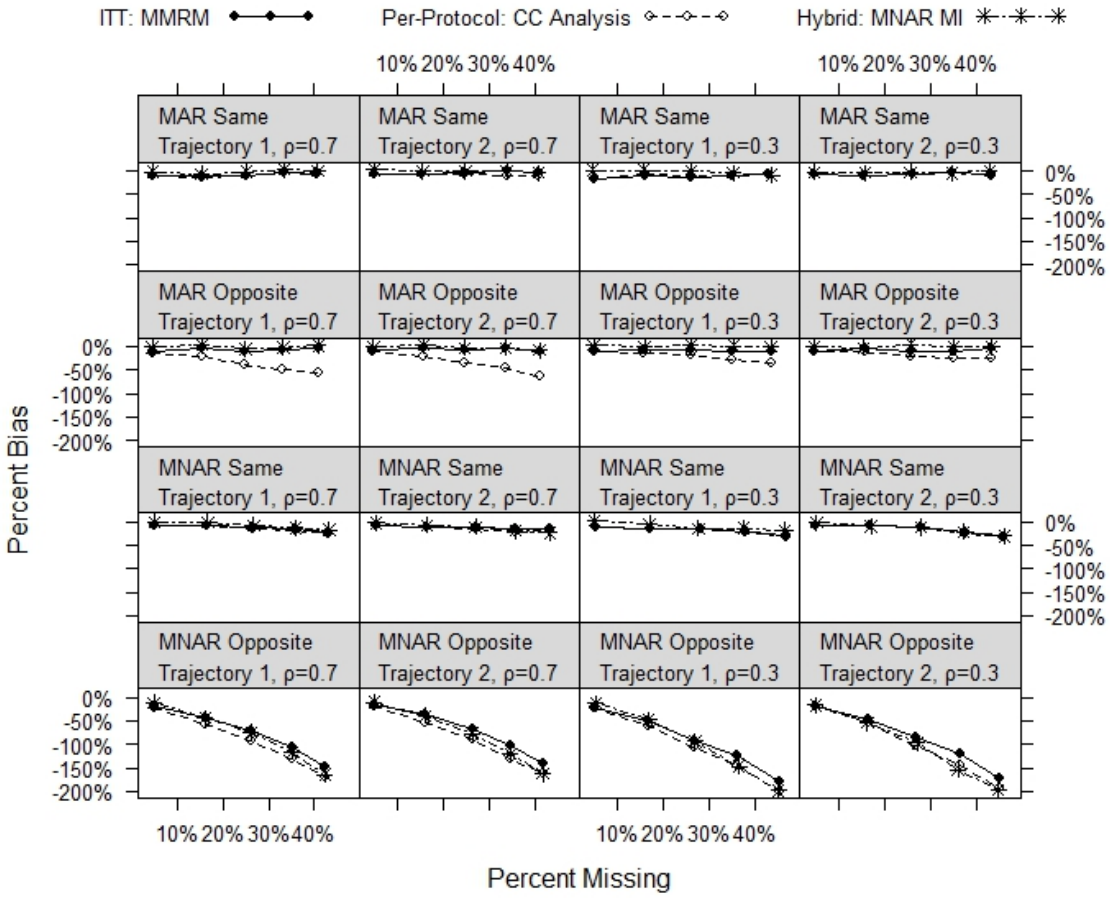

Figure 4.1: Comparison of Intention-to-Treat and Per-Protocol Analysis with Percent Bias vs. Percent Missing Data Curves. Intention-to-treat set is analyzed with a mixed model for repeated measures (ITT: MMRM), per-protocol analyzes complete cases with 2-sample t-test (Per-protocol: CC Analysis), and hybrid approach uses multiple imputation with an MNAR control-based model (Hybrid: MNAR MI). Missing proportions range from 5% to 50% under different missingness mechanisms: MAR or MNAR, same or opposite direction of missingness, early or late separation of trajectory of effects (trajectory 1 and trajectory 2 respectively), and $\rho = 0.3$ or $\rho = 0.7$. Subject compliance rate is approximately 80%.

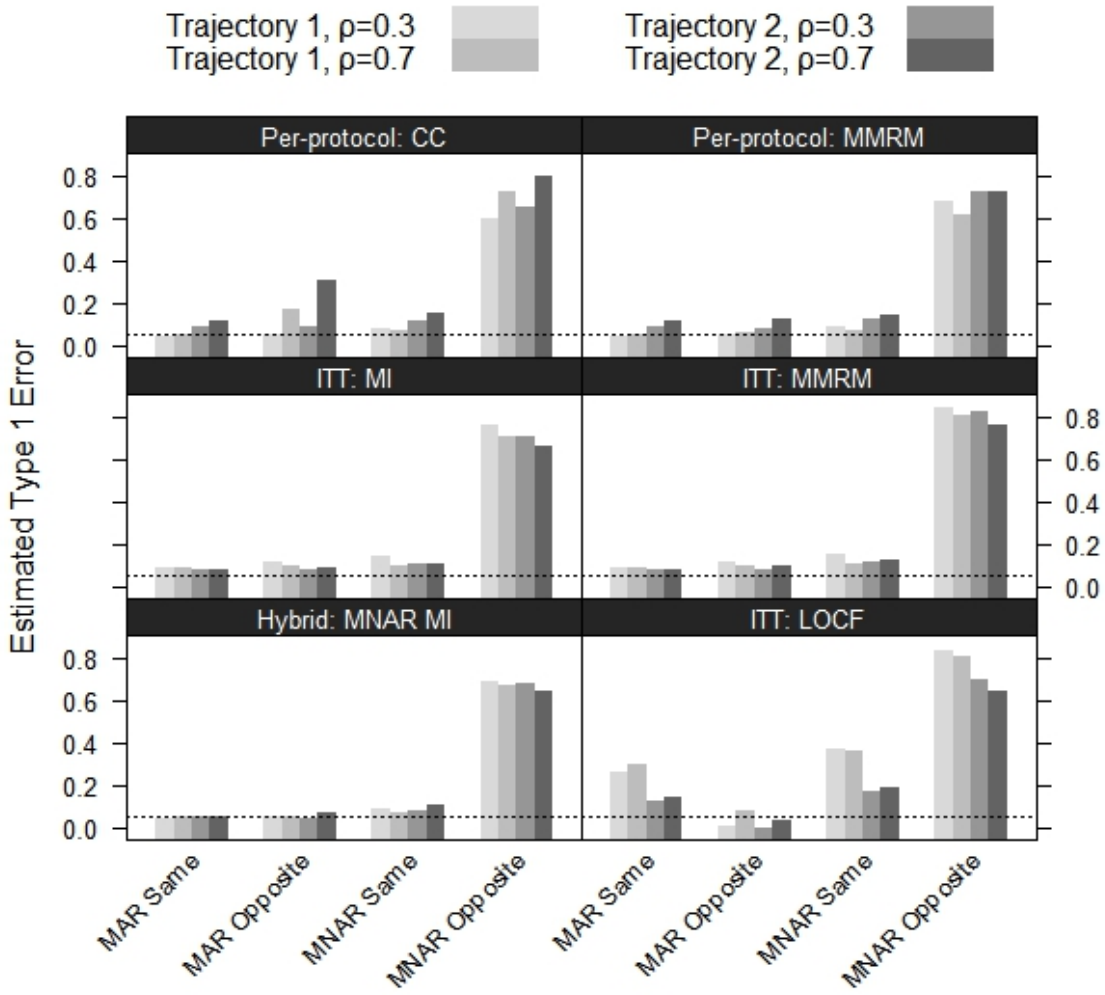

Figure 4.2: Estimated type I error under a variety of missing data patterns and data structures, MAR-same, MAR-opposite, MNAR-same, MNAR-opposite, under both patterns of effect trajectories: trajectory 1 (early separation); trajectory 2 (late separation). Percent of missing data is fixed at 20%. The dashed horizontal lines are drawn at 0.05, the presumed type 1 error rate.

Simulations under the alternative hypothesis, where the trajectories of effects in treatment arms were equivalent, are reported in Table 4.3. Power was highest for those approaches whose corresponding type 1 error rates were also high, showing consistent anti-conservativism. When the data were MNAR in opposite directions, all the approaches greatly underestimated the true difference between arms. Aside from the approaches using last-observation-carried-forward and the intention-to-treat/per-protocol hybrid, power rates were above 90%. With low correlation between repeated measures, the hybrid multiple imputation approach maintained reasonable power rates ranging from 81% to 95% when data were MAR or with no missing data. Power estimates for simulations of full compliance are reported in the supplemental materials.

4.4 Example: Letrozole Study

We demonstrate the six analysis approaches on real data from a four-arm non-inferiority trial comparing dosing regimens of letrozole, an estrogen suppressing drug used for treatment of hormone sensitive breast cancer[38]. A sample of 112 healthy, post-menopausal women with increased cancer risk were randomized to receive letrozole at 2.5, 1.0 or 0.25 mg three times a week (experimental arms) or 2.5 mg daily (active control arm). The primary endpoint was percentage of serum estradiol suppression after 24 weeks of treatment. Treatments were non-inferior if the mean percentage of serum estradiol suppression was at least 70% of the percent suppression in the active control. In this example, for simplicity we limit our analysis to comparison of change from baseline serum estradiol at week 24 in the active control and the 0.25 mg treatment arm using a non-inferiority margin of -2.5 estradiol pg/mL. Because lower values of serum estradiol are favorable the estimated difference is non-inferior if the lower limit of the 90% confidence interval, L, for $\mu_{AC} - \mu_T$ is $L > -2.5$. Subjects were considered compliant if they were at least 95% compliant with the treatment regimen, and 24 week measurements were set to missing if subjects were recorded as non-completers.

There were 28 participants randomized to each of the treatment arms. At baseline, the mean estradiol (mg/dL) was 6.08 (SD=6.13) and 5.35 (2.77) in the T and AC arms

Table 4.3: Bias and Power with 80% Compliance and 20% Missing Data. Estimates for change from baseline difference and power under the alternative hypothesis of non-inferiority where the true difference is 0. A: Last Observation Carried Forward. B: Mixed Model for Repeated Measures. C: Multiple Imputation. D: Multiple Imputation with MNAR Control-Pattern. E. Bias (estimate) for $\mu_A C - \mu_T = 0$. F. Percent power.

	Intention to Treat					
	LOCF[A]		**MRMR[B]**		**MI[C]**	
	Bias[E]	Power[F]	Bias	Power	Bias	Power
Moderate Correlation $\rho = 0.7$						
No Missing	-	-	0.010	99.7	-	-
MAR - Same	0.003	99.5	0.001	99.0	-0.011	99.4
MAR - Opposite	0.205	97.4	-0.001	98.9	-0.001	98.6
MNAR - Same	-0.006	97.5	-0.005	99.2	-0.01	98.9
MNAR - Opposite	-0.268	100.0	-0.534	100.0	-0.448	100.0
Low Correlation $\rho = 0.3$						
No Missing	-	-	-0.007	98.5	-	-
MAR - Same	0.008	97.2	0.006	95.9	0.005	91.3
MAR - Opposite	0.541	53.8	-0.006	94.9	-0.007	91.2
MNAR - Same	-0.012	98.1	-0.01	96.6	-0.008	91.8
MNAR - Opposite	-0.392	100.0	-0.687	100.0	-0.582	100.0

	Per Protocol				**Hybrid**	
	Complete Cases		**MMRM**		**MI - MNAR[D]**	
	Bias	Power	Bias	Power	Bias	Power
Moderate Correlation $\rho = 0.7$						
No Missing	-	-	0.012	99.2	0.002	99.0
MAR - Same	-0.005	97.3	-0.005	97.5	-0.015	98.7
MAR - Opposite	-0.184	99.8	-0.001	98.0	-0.007	97.4
MNAR - Same	-0.005	98.6	-0.007	98.6	-0.01	98.7
MNAR - Opposite	-0.604	100.0	-0.511	100.0	-0.546	100.0
Low Correlation $\rho = 0.3$						
No Missing	-	-	-0.01	95.1	-0.019	94.6
MAR - Same	0.007	90.1	0.007	90.6	0.009	83.2
MAR - Opposite	-0.032	91.2	-0.006	90.4	0.002	81.8
MNAR - Same	-0.008	91.8	-0.008	92.0	-0.009	84.9
MNAR - Opposite	-0.652	99.9	-0.692	100.0	-0.721	99.9

Table 4.4: Estimates and 90% Confidence Intervals for the Difference in Change From Baseline Serum Estradiol (pg/mL)

Intention to Treat			**Per Protocol**		**Hybrid**
LOCF	**MMRM**	**MI**	**Complete Cases**	**MMRM**	**MI - MNAR Control**
0.42 (-1.90, 2.70)	0.68 (-1.58, 2.94)	0.71 (-1.51, 2.93)	1.23 (-1.44, 3.89)	0.71 (-1.56, 2.97)	0.68 (-1.56, 2.91)

respectively. At week 24, three subjects from the T arm and one subject from AC did not complete the treatment and their estradiol values were considered missing. Five subjects were less than 95% compliant with treatment protocol (3 in T, 2 in AC) and were modeled as non-compliant subjects. At 24 weeks among completers, means reduced to 1.00 (1.58) for T and 0.80 (1.23) for AC. Table 4.4 shows estimates and 90% confidence intervals for the letrozole study data for each of the six methods. The estimated difference in change from baseline estradiol between the treatment and control arms were positive for all methods, indicative of better performance in the T arm. All lower 90% confidence limits were greater than -2.5 and showed non-inferiority.

4.5 Discussion

The objectives of this simulation study were to examine bias, type 1 error rates, and power in a two-arm longitudinal non-inferiority trial under several missing data scenarios. By varying features of the data, such as the trajectory of effects and degree of correlation between the repeated measures, we aimed to identify robust analytic approaches that could be used to inform selection of primary estimands in future trials. Our results showed that there is potential for substantially biased estimates when data are MNAR in opposite directions. Other missing data mechanisms were controlled well with a likelihood-based approach (mixed model for repeated measures), which overall outperformed both complete case analysis and last-observation-carried-forward imputations. Use of compliance status as an auxiliary variable in a multiple imputation model was effective in a hybrid intention-to-treat/per-protocol analysis that removed and imputed values from non-compliant subjects: this analysis proved to be the most conservative and gave unbiased estimates with minimal reduction in power. In our real data example, the difference in point estimates between the last-observation-carried-forward, complete cases and mixed model analyses suggest that conclusions may be sensitive to different assumptions about the missing mechanism, and that estimates may differ substantially according to the definition of the analysis set, even with relatively small amounts of missing data or non-compliance.

The predominant cause of bias and inflated type 1 error in both intention-to-treat and per-protocol analyses was a MNAR mechanism with treatment arm interaction (missingness in opposite directions). Under this missing data scenario, even if the true effect difference between arms exceeded the non-inferiority margin, not only might a conclusion of non-inferiority be incorrectly found, but an ineffective treatment could be found superior to the active control if an analysis plan called for sequential testing in the manner of (1) test non-inferiority (2) if non-inferior then test superiority[2]. Even if the true treatment difference were 0 (so that a non-inferior conclusion were correct), an analytic approach that fails to account for this missing data mechanism will produce such biased estimates that the effectiveness of the experimental treatment might be radically overstated. While such a MNAR mechanism may be rare in practice, it is worthwhile for clinicians and analysts alike to be aware of this worst-case scenario. New approaches to sensitivity analyses should reflect vulnerabilities of the non-inferiority design to this type of missingness pattern.

As others have shown via both simulation and theoretical properties, last-observation-carried-forward imputation can introduce substantial bias if the underlying assumptions are violated[21,22]. In superiority trials, the use of last-observation-carried-forward might be justified on grounds that a biased result is more likely to be conservative[23]. However, we have shown last-observation-carried-forward imputation can be conservative or anti-conservative depending on the dropout mechanism and the trajectory of effects, which are generally unknown. Furthermore, a tendency for conservativism in superiority studies may translate to anti-conservativism in non-inferiority trials. In our simulations with data MAR in opposite directions, high correlation between repeated measures highlighted the inadequacy of complete case analysis. Overall, the mixed model for repeated measured nearly always outperformed complete case analysis and last-observation-carried-forward; hence, these approaches should be used only if known information about the missing mechanism or non-compliance present strong rationale for doing so.

Our simulations agree with findings of Sanchez and Chen[21]: the intention-to-treat and per-protocol analyses may be biased in the same direction. On the other hand, under certain conditions, the per-protocol analysis may be anti-conservative result and the intention-to-

treat may be conservative. Hence, it may not be necessary nor desirable to give equal weight to these analyses, as has been recommended by some regulatory agencies[12,24].

Sanchez and Chen[21] proposed the use of a hybrid analysis set to reduce bias from protocol violations and while addressing missing data in a statistically principled manner. The hybrid intention-to-treat/per-protocol approach we developed retains the benefits of randomization while accounting for post-randomization protocol violations associated with both missingness and outcomes. The hybrid approach may conform well with what the International Conference on Harmonisation[6] E9 describes as a hypothetical strategy, whereby the effort is to measure treatment differences between those who would adhere to protocol (i.e. with full protocol compliance.) Such an analysis would likely have to be predefined and consistent with the trials primary estimand to be judged acceptable. However, our results show that, while reasons for non-compliance should themselves be examined and understood, in a non-inferiority context, an estimand defined to measure differences between hypothetical compliers may indeed be the conservative one. If a hypothetical estimand cannot be justified, trialists may choose an approach aligned with the intention-to-treat principle, as in the "Treatment Policy" strategy[6]. Our results provide evidence that a mixed model, perhaps implemented within a multiple imputation framework, has the advantage by better controlling type 1 error than other imputation approaches. With the inclusion of meaningful auxiliary variables associated with outcome or probability of missingness, the multiple imputation process may correct for underestimation of uncertainty in the mixed model estimates.

4.5.1 Strengths and Limitations

We should be cautious to generalize these findings given the natural limitations of simulation studies. It is not possible to test every scenario however common or rare. Still, along with the works of Wiens and Rosenkranz[5] and Yoo[22] we have produced evidence to support the use of likelihood-based and multiple imputation approaches and insight into various missing mechanisms in non-inferiority trials. Simulation studies typically investigate missing data mechanisms separately; however, it is likely that in real trials the missing data are caused

by a variety of mechanisms, some of which are known and others that are unknown. At the least, this study provides a benchmark for assessing potential influence of missing values.

The hybrid intention-to-treat/per-protocol multiple imputation analysis we describe may not be appropriate in data with certain types of protocol violations. Although protocol violations can take many forms including, but not restricted to, use of an alternative treatment, rescue medication, treatment switching, or death, we chose to focus on the simple case of non-compliance due to inefficacy of treatment. Other protocol violations, such as non-compliance due to toxicity, may be a meaningful endpoint in and of itself. The International Conference on Harmonisation [6] E9 makes clear that "A scientific question of interest based on the effect if all subjects had adhered to treatment is not well-defined without a thorough discussion of the hypothetical conditions under which it is supposed that they would have adhered" [6]. Thus, trialists must understand the full nature of protocol violations in their data before a hypothetical analysis can be undertaken and meaningfully interpreted.

4.6 Conclusion and Recommendations

The usefulness of any missing data tool is highly situational, and care must be taken to thoughtfully apply methods with consideration of the study design, relationships between the variables, what is known about the missing data, and the research questions. We do not advise that researchers adopt a blanket approach to missing data handling. While no uniformly best method exists for analyzing missing data, in longitudinal non-inferiority trials we recommend the mixed model for repeated measures with unstructured time and unstructured covariance if the number of time points and data collected allow for it. Furthermore, the use of auxiliary data within a multiple imputation framework can enhance the accuracy of results and better control type 1 error rates by accounting for uncertainty inherent in the imputation process.

Ultimately, the choice of analysis set, missing data handling and analysis procedure follows from the choice of estimand, which is principally made in the design stage. Estimands that attempt to measure effects in the fully compliant sample may be less susceptible to type 1 error and be more interpretable in the non-inferiority setting. In any case, it is particularly

important that the active control perform as expected based on historic trials so that a conclusion of non-inferiority is evidence of superiority of the treatment versus placebo. It is incumbent upon clinicians and trial designers to understand the correct interpretations of their predefined estimands, and to be cognizant of the estimands' susceptibilities to bias. Knowledge of how various trial characteristics affect estimates is therefore critical to the selection of an estimand and its interpretation.

If the trial's primary estimand is concerned with estimating treatment effects in a fully compliant sample, then our hybrid MNAR control-based multiple imputation approach is viable. Otherwise, the approach can be used as complementary to a pre-defined analysis that may have failed to anticipate intercurrent events. While missing data have been studied extensively, since the National Research Council report[39], there is some evidence that researchers are improving efforts to avoid missing data altogether. Yet, intercurrent events will continue to be problematic for statistical analysis.

We recommend that last-observation-carried-forward imputation be avoided in non-inferiority trials as it can heavily influence estimates, increasing bias most often in the direction of the alternative hypothesis. Similarly, complete case analysis, while simple to implement, is not advised unless there is a strong justification for doing so. Some authors have argued that multiple imputation, while generally a principled approach to missing data, may be unnecessary and inefficient by introducing Monte Carlo error[25]; however, we view the conservatism of multiple imputation in non-inferiority trials as an added benefit and hedge against the anti-conservatism of other intention-to-treat approaches. When one does not know the missing data mechanism, a conservative approach should include assessing the robustness of conclusions while assuming a worst-case scenario where data are MNAR in opposite directions. Further research is needed to explore how best to implement and draw conclusions from sensitivity analyses with respect to missing data assumptions for non-inferiority trials.

This work is currently in revision with *Statistics in Biopharmaceutical Research*.

Chapter 5

Sensitivity Models for Informative Dropout in Non-Inferiority Trials

Abstract

A pattern-mixture model approach is proposed for conducting sensitivity analysis with respect to missing data assumptions in a non-inferiority trial with longitudinal continuous outcomes. The model is fit in a multiple imputation framework using identifying restrictions equating monotone dropout patterns to the observed patterns for intermediary time points. At the final time point, the missing data are imputed from the fully observed patterns plus a shift parameter that differs by treatment arm. We allow the shift parameters to vary from $-\delta/2$ to $\delta/2$, where δ is the pre-defined non-inferiority margin. Next, we present an alternate approach for imputing data in the active control arm using an informative prior on the distribution of outcomes. With a Markov Chain Monte Carlo model, we multiply impute missing data from the predictive posterior distribution based on a multivariate normal prior with hyperparameters that could be obtained by meta-analysis of historical investigations of the active control. We demonstrate two applications of our approach in an enhanced tipping point analysis with graphical display. Finally, we conduct simulations to compare a linear mixed model with several versions our approach for bias, type I error, power under various missing data mechanisms.

5.1 Introduction

In non-inferiority clinical trials, which are more sensitive to protocol deviations and by nature less conservative than superiority studies, the statistical challenges presented by missing outcomes warrant special attention. Among the challenges is that non-compliance, dropout, or other post-randomization events can lead to anti-conservative conclusions[12,21,33]. Subjects who deviate from protocol may be similar: they may be more susceptible to adverse events or be in worse condition than average. Thus, if the outcomes are measured and analyzed on the full sample including the non-adherent, the estimated difference in effect between treatment arms may be attenuated. On the other hand, if outcomes are simply unobserved, the analyst must either exclude missing cases or assume a mechanism that explains subject dropout. In either case, the analyses may fail to control type I error rates and risk incorrectly concluding non-inferiority.

Sensitivity analyses explore the dependence of conclusions from the primary analysis on assumptions about missing data and possibly other protocol deviations. These assumptions typically define the nature of associations between the missing values, the probabilities of missingness and other variables, either observed or unobserved. A well-designed sensitivity analysis will first make efforts to condition imputation upon observed variables suspected to be associated with the missing outcomes (effectively making the data "more Missing at Random (MAR)"[40]. Then, the analyst can explore plausible scenarios with an imputation model of informative dropout which assumes the mechanism is Missing Not at Random (MNAR). One can argue MNAR models lack robustness because nothing in the observed data give us any distributional information about the relationships between covariates (or outcomes observed at earlier timepoints) and the missing outcome values[41]. However, together with the Bayesian nature of the multiple imputation procedure, historic trial data on the active control arm offer an alternative. Historical borrowing can improve trial efficiency and may justify the usage of strong priors for estimation of a Bayesian imputation model in sensitivity analysis.

Conservative missing data imputation for superiority trials may include imputation models that assume a similar (or the same) mechanism governing missing data across arms.

Given the potential for anti-conservatism of non-inferiority trials, a sensitivity analysis that imputes separately by treatment arm is warranted. Some authors have proposed imputing under the assumption of the null hypothesis[2,5]. One such method uses a last-observation-carried-forward (LOCF) imputation with a shift by δ, the non-inferiority margin, either upwards or downwards according to the treatment arm assignment[12].This approach is highly susceptible to bias; the LOCF approach has been repeatedly shown to be unreasonable in most situations[18].

In this chapter, we aimed to develop a robust sensitivity analysis that will identify MNAR possibilities that raise doubts about a false positive but that will support a correct primary analysis under the alternative. We present a pattern-mixture model approach for imputing informative dropout in non-inferiority trials with continuous outcomes[42]. First, we present a general pattern-mixture model for the joint distribution of the missing mechanism and outcomes. Next, we present identifying restrictions and fit a multiple imputation model for the missing values. We extend this approach by imputing separately by treatment arm and use an informative prior to estimate the posterior predictive distribution for the outcomes in the active control. Each method is implemented as a tipping point analysis where a shift parameter alters the multiply imputed data by treatment arm systematically. With Monte Carlo simulations, we compare the performance of three implementations of our approach with a primary analysis using a linear mixed model. Finally, we demonstrate our models application as a tipping point analysis with a visualization of selected simulated data sets.

5.2 Pattern-Mixture Models for Informative Dropout

5.2.1 Missing Data Mechanisms

We formally define the categories of missing data mechanisms as first enunciated by Rubin[3] in the context of a longitudinal trial with a monotone missing data pattern consistent with dropout. Let $\mathbf{Y} = (Y_1, ...Y_T)$ be the vector of outcomes measured T times. Let $\mathbf{X}$ be a vector of covariates that includes the treatment assignment. Let the missing data indicator, R, be defined as the time of last observation, so that $R = T$ means that all times are

observed. Let ψ be an unknown parameter governing the distribution of the missing data mechanism and let θ be the parameter of scientific interest. The data are said to be Missing Completely at Random (MCAR) if

$$f(R|\mathbf{X}, \mathbf{Y}, \theta, \psi) = f(R|\psi),$$

Missing at Random (MAR) if

$$f(R = t|\mathbf{X}, \mathbf{Y}, \theta, \psi) = f(R = t|\mathbf{X}, Y_1, ..., Y_t, \psi),$$

and Missing not at Random (MNAR) whenever

$$f(R = t|\mathbf{X}, \mathbf{Y}, \theta, \psi) = f(R = t|Y_1, ..., Y_{t+1}, ..., Y_T, \theta, \psi).$$

In trials, data MCAR are, on average, distributed evenly across treatment arms, are unrelated to any variables under study, and have little effect on estimates; however, they may reduce statistical power.

Under MAR, inferences about θ can be based on the likelihood $L(\theta|\mathbf{Y_{obs}})$ where $\mathbf{Y_{obs}}$ represents the set of observed values of $\mathbf{Y}$, and $\mathbf{Y_{mis}}$ represents the missing values of $\mathbf{Y}$, and $\mathbf{Y} = \{\mathbf{Y_{obs}}, \mathbf{Y_{mis}}\}$. To see this, consider the joint distribution $f(\mathbf{Y}, R|\mathbf{X}, \theta, \psi) = f(\mathbf{Y}|\mathbf{X}, \theta) \cdot f(R|\mathbf{X}, \mathbf{Y}, \psi)$. Then, integrating over $\mathbf{Y_{mis}}$ leads to

$$\begin{aligned} f(\mathbf{Y_{obs}}, R|\mathbf{X}, \theta, \psi) &= \int f(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{X}, \theta) \cdot f(R|\mathbf{X}, \mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \psi) d\mathbf{Y_{mis}} \\ &= f(R|\mathbf{X}, \mathbf{Y_{obs}}, \psi) \int f(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{X}, \theta) d\mathbf{Y_{mis}} \text{ (by MAR property)} \\ &\propto \int f(\mathbf{Y_{obs}}, \mathbf{Y_{mis}}|\mathbf{X}, \theta) d\mathbf{Y_{mis}} \\ &= f(\mathbf{Y_{obs}}|\mathbf{X}, \theta). \end{aligned}$$

Provided the parameters spaces of θ and ψ are distinct, under MAR, the missing mechanism is said to be ignorable. Non-ignorable missing data generally refer to the MNAR case when $P(R = t)$ depends at minimum on the value of Y_{t+1}.

5.2.2 Pattern-Mixture Models

Models proposed for likelihood-based approaches to handling non-ignorable dropout are typically either pattern-mixture models or selection models[43]. Each represents a different factorization of the joint distribution of outcomes and missing data indicator, $(\mathbf{Y}, R)$ and in practice, lead to a different set of simplifying assumptions[42]. The pattern-mixture model factorizes the distribution as

$$f(\mathbf{Y}, R|\mathbf{X}, \theta, \psi) = f_R(R|\mathbf{X}, \psi) f_{Y|R}(\mathbf{Y}|\mathbf{X}, R, \theta),$$

whereas selection models are factorized as

$$f(\mathbf{Y}, R|\mathbf{X}, \theta, \psi) = f_Y(\mathbf{Y}|\mathbf{X}, \theta) f_{R|Y}(R|\mathbf{X}, \mathbf{Y}, \psi).$$

Molenberghs et al.[42] showed that while the missing data mechanisms described by Rubin are more easily interpretable within the selection model framework, they can be described equivalently by pattern mixture model formulations.

Modeling informative dropout (i.e. data MNAR) requires distributional assumptions about the missing data that cannot be verified (since the data are unobserved). Thus, direct estimation of θ can be susceptible to bias. Instead, we can employ the pattern-mixture model framework for imputing missing outcome data using multiple imputation in a sensitivity analysis. This has several advantages, among them is that the analysis model remains consistent across sensitivity and primary analyses, and the imputation model can be based on a different model which conditions on variables not of primary interest. Additionally, we can benefit from the interpretability of the pattern-mixture formulation and its flexibility in defining missing data assumptions. Furthermore, it is easily implemented in readily available statistical software.

Expanding the joint distribution of the pattern-mixture formulation into the missing and observed outcomes gives

$$f(\mathbf{Y_{mis}}, \mathbf{Y_{obs}}, R|\mathbf{X}, \theta, \psi) = f(\mathbf{Y_{obs}}, R|\mathbf{X}, \theta, \psi) f(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, \mathbf{X}, R, \theta, \psi).$$

If the data are MAR, then

$$f(R|\mathbf{Y}, \mathbf{X}, \psi) = f(R|\mathbf{Y_{obs}}, \mathbf{X}, \psi).$$

This implies the distribution of missing data, conditional upon the observed data, is independent of R:

$$\begin{aligned} f(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, R, \mathbf{X}, \theta, \psi) &= \frac{f(R, \mathbf{Y}|\mathbf{X}, \psi) f(\mathbf{Y}|\mathbf{X}, \theta)}{f(\mathbf{Y_{obs}}, R|\mathbf{X}, \theta, \psi)} \\ &= \frac{f(R|\mathbf{Y_{mis}}, \mathbf{Y_{obs}}, \mathbf{X}, \psi) f(\mathbf{Y_{mis}}, \mathbf{Y_{obs}}|\mathbf{X}, \theta)}{f(\mathbf{Y_{obs}}, R|\mathbf{X}, \theta, \psi)} \\ &= \frac{f(R|\mathbf{Y_{obs}}, \mathbf{X}, \psi) f(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, \mathbf{X}, \theta) f(\mathbf{Y_{obs}}|\mathbf{X}, \theta)}{f(\mathbf{Y_{obs}}, R|\mathbf{X}, \theta, \psi)} \\ &= f(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, \mathbf{X}, \theta). \end{aligned}$$

However, when data are MNAR, for arbitrary dropout patterns,

$$f(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, \mathbf{X}, R = q) \neq f(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, \mathbf{X}, R = r),\ q \neq r.$$

Thus, the distributional assumptions about variables with missing data cannot be based soley upon cases with fully observed values. However, the formulation allows the modeler to impute from the observed distribution and systematically vary the imputed values to reflect the degree to which the distributions of observed and unobserved values differ. The advantage of using the pattern-mixture model approach is that we do not have to explicitly fit an MNAR model, rather, we can impute values under the MAR hypothesis and postulate a mean or scale shift among the missing data. With this approach, analysts can easily communicate alternative assumptions for sensitivity analysis [44].

5.2.3 Application to a Longitudinal Non-Inferiority Trial

We present a pattern-mixture model for generating multiple imputations in a two-arm non-inferiority trial, with experimental treatment arm denoted ET and active control denoted by AC. For simplicity, we restrict the number of data collection times to 3 (baseline and

two subsequent measurements); an extension to any arbitrary number of timepoints is straightforward.

Analysis Model

We perform a primary analysis of the change from baseline difference between treatment arms at the final time using the so-called mixed model for repeated measures (MMRM), where a length 3 vector of outcomes, $\mathbf{Y}$ is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim N(0, \mathbf{V}).$$

$\mathbf{X}$ is a $3 \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effects, $\boldsymbol{\epsilon}$ is the 3×1 vector of residuals, and $\mathbf{V}$ is the 3×3 unstructured variance-covariance matrix for $\mathbf{Y}$. $\mathbf{V}$ is estimated from the data and allowed to be unstructured to accommodate correlations between the repeated measurements on subjects. Although this model can require a large number of parameter estimates if the design includes many timepoints, the model is not vulnerable to misspecification if the response profiles are non-linear and is robust to data MAR[18].

The fixed effects include the standard intercept, categorical time (hence, there are no constraints on the means), treatment arm indicator and the time-treatment interaction, hence, $p = 6$. In scalar form, the analysis model estimates the expected outcome for a given subject, i, at time j, $(j = 0, 1, 2)$, by:

$$E[Y_{ij}] = \beta_0 + \beta_1 treat_i + \beta_2 time1_j + \beta_3 time2_j + \beta_4 treat_i \times times1_j + \beta_5 treat_i \times time2_j,$$

where

$treat_i = 1$ if subject i is in the ET arm and $treat_i = 0$ otherwise;

$time1_j = 1$ if $j = 1$ and $time1_j = 0$ otherwise;

$time2_j = 1$ if $j = 2$ and $time2_j = 0$ otherwise.

The mean treatment effects for treatment arm $s \in \{AC, ET\}$ at time $j \in \{0, 1, 2\}$, denoted by μ_{st}, are linear combinations of the elements of $\boldsymbol{\beta}$:

$$\mu_{AC,0} = \beta_0$$
$$\mu_{AC,1} = \beta_0 + \beta_2$$
$$\mu_{AC,2} = \beta_0 + \beta_3$$
$$\mu_{ET,0} = \beta_0 + \beta_1$$
$$\mu_{ET,1} = \beta_0 + \beta_1 + \beta_2 + \beta_4$$
$$\mu_{ET,2} = \beta_0 + \beta_1 + \beta_3 + \beta_5$$

Without loss of generality, we assume that larger values of μ_{st} are desirable. The primary outcome of scientific interest is the difference between arms in the mean change from baseline at the final time point, $\Delta = (\mu_{AC,2} - \mu_{AC,0}) - (\mu_{ET,2} - \mu_{ET,0}) = \beta_5$. Controlling type I error at 5%, non-inferiority of ET is established if the upper 90% confidence limit on Δ is less than δ, the non-inferiority margin. We consider only missing data in a monotone

Table 5.1: Means of outcomes in treatment arm s, $s = \{AC, ET\}$, at each time and for every pattern of monotone missing data with all baseline values observed. The parameters in gray are inestimable without additional assumptions.

	Time		
Pattern, k	t_0	t_1	t_2
0	$\mu_{s,0}^{(0)}$	$\mu_{s,1}^{(0)}$	$\mu_{s,2}^{(0)}$
1	$\mu_{s,0}^{(1)}$	$\mu_{s,1}^{(1)}$	$\mu_{s,2}^{(1)}$
2	$\mu_{s,0}^{(2)}$	$\mu_{s,1}^{(2)}$	$\mu_{s,2}^{(2)}$

dropout pattern and require baseline observations; hence, in this context there are three dropout patterns: $k = 0$ for completely observed subjects, $k = 1$ for data with one missing observation, and $k = 2$ for data with only baseline values, as shown in Table 5.1

Identifying Restrictions

The key feature of the pattern-mixture model approach is that we do not assume the same underlying distribution for cases with different missing data patterns. Instead, we introduce a conservative bias to our model for missing values toward the null hypothesis of inferiority.

To achieve this, we applied the following constraints:

$$\mu^{(2)}_{ET,2} = \mu^{(1)}_{ET,2} = \mu^{(0)}_{ET,2} - \delta/2 \tag{5.1}$$

$$\mu^{(2)}_{AC,2} = \mu^{(1)}_{AC,2} = \mu^{(0)}_{AC,2} + \delta/2 \tag{5.2}$$

$$\mu^{(0)}_{s,1} = \mu^{(1)}_{s,1} = \mu^{(2)}_{s,1} \text{ for } s \in \{ET, AC\} \tag{5.3}$$

Equation 5.1 models the missing values from the ET arm by matching missing patterns to the complete data; however, the expected mean is shifted downward by half the non-inferiority margin for subjects with dropout. Equation 5.2 shifts the treatment effect for dropouts in the AC arm upwards by $\delta/2$ compared with their counterparts with complete data. Data missing from intermediary timepoints are assumed to be from the same distribution as the observed data, represented by equation 5.3.

Together, these constraints impose an MAR assumption on data missing before the final timepoint, and they assume that at the final observation period, the missing values themselves are related both to the probability of dropout and to the treatment arm assignment. The result is that, if the complete data are similar at the final timepoint ($\mu^{(k)}_{AC,2} = \mu^{(k)}_{ET,2}$ for $k = 0$), then for missing patterns $k \in \{1, 2\}$, $\mu^{(k)}_{AC,2} - \mu^{(k)}_{ET,2} = \delta$. Specifically, the constraints model a worst-case scenario wherein subjects assigned to ET with worse outcomes are more likely to drop out, whereas subjects assigned to AC are more likely to drop out with positive outcomes. The likelihood of such a mechanism can be evaluated by eliciting expert opinion, using available information about the impact of other variables such as side effects, adverse events, and efficacy of the active control[45].

5.2.4 Fitting the Pattern-Mixture Model with Multiple Imputation

Multiple imputation is a procedure that involves three steps:

1. Fill in the missing observations with imputed values m times. This is achieved by fitting an imputation model.

2. Fit the primary analysis model to the completed data sets to obtain m sets of estimates, $\hat{Q}$, and standard errors, $\hat{W}$, for Q, the parameter of interest.

3. Combine the m estimates and standard errors according to rules proposed by Rubin [46]:

$$\begin{aligned}\text{Point estimate: } \overline{Q} =& \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \\ \text{Variance: } T =& \overline{W} + \Big(1+\frac{1}{m}\Big)B \\ =& \frac{1}{m}\sum_{i=1}^{m}\hat{W}_i + \Big(1+\frac{1}{m}\Big)\frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})^2.\end{aligned}$$

The pooled estimate is simply the arithmetic mean of the m estimates; however, the pooled standard error takes account of the uncertainty both due to sampling variability and due to the missing observations. It can be shown that $\frac{Q-\bar{Q}}{\sqrt{T}}$ follows a t-distribution with degrees of freedom depending on m [35].

The pattern-mixture model we have proposed for sensitivity analysis is easily implemented within a multiple imputation framework. One of the advantages of using multiple imputation for sensitivity analysis is that we may use variables in the imputation model that are related to missing data but that are not of interest in the primary analysis. As such, the sensitivity analysis with an MNAR imputation means the analysis model remains consistent with that of the primary analysis.

In the next sections, we introduce two methods of multiple imputation that are implemented in existing software [37]. We will then show how imputed values from these methods can be systematically shifted to satisfy the identifying restrictions (Eqs. **??**) of our MNAR pattern-mixture model.

A Linear Regression Imputation Model

The regression method [47] for multiple imputation assumes the data are normally distributed and that the missing data pattern is monotone. Each multiply imputed data set is generated by imputing missing values across the variables one at a time. Specifically, to generate one multiply imputed data set, we use the data with observed to fit the following regression

model:

$$Y_{i,1} = \beta_0 + \beta_1 treat_i + \beta_2 Y_{i,0} + \epsilon,\ \epsilon \sim N(0, \sigma^2).$$

Given the estimated regression coefficients, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and covariance matrix $\sigma^2 X'X$, where X is the design matrix derived from the intercept, treatment indicator, $treat_i$ and $Y_{i,0}$, random draws from the coefficient and variances distributions, $\beta_{*0}, \beta_{*1}, \beta_{*2}, \sigma_*$ are used to replace the missing values in $Y_{i,1}$ by

$$\beta_{*0} + \beta_{*1} treat_i + \beta_{*2} Y_{i,0} + z\sigma_*,$$

where z is a draw from a standard normal. The process is repeated for $Y_{i,2}$, where the regression model is fit with the observed data:

$$Y_{i,2} = \beta_0 + \beta_1 treat_i + \beta_2 Y_{i,0} + \beta_3 Y_{i,1} + \epsilon,\ \epsilon \sim N(0, \sigma^2).$$

Subsequent imputed datasets are generated from repeated draws of the distribution of β covariates and σ.

A Bayesian Imputation Model with Informative Priors

The Markov Chain Monte Carlo (MCMC) method is used to approximate probability distributions that are analytically intractable[47]. The probability distribution of interest is generated from a sequence of random variables, a Markov Chain, where each random variable in the sequence depends only on the previous random variable. After many steps of the Markov chain, repeated steps are approximate random draws of the stationary distribution. For a full discussion of MCMC methods, see Kroese et al.[48].

MCMC methods have been used to approximate posterior distributions in inferential Bayesian statistics. In missing data problems, we can simulate the complete-data posterior distribution $p(\theta|\mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \mathbf{X})$ by augmenting the observed values with simulated or estimated missing values. The data are assumed multivariate normal.

For the sensitivity analysis of non-inferiority trials, we sort data by treatment arm and generate the full joint distribution of variables for each group separately. Hence, for each

treatment arm, the imputed values for the missing data are gotten from the estimated posterior distribution:

$$p(\theta|\mathbf{Y_{obs}}, \mathbf{Y_{mis(t=1)}}, \mathbf{Y_{mis(t=2)}})$$

Note that $\mathbf{Y_{mis(t=0)}}$ does not appear in the distribution because we have required that baseline values be observed; however, the MCMC method does not require a monotone missing data pattern.

The posterior is estimated by a series of repeating steps: the imputation (I) step and the posterior (P) step.

The I-step: For an estimated mean vector and covariance matrix, we impute the missing values for each subject independently. The imputations are random draws from the estimated distribution of missing variables conditioned on the observed, i.e. $\mathbf{Y_{mis(t=1)}}, \mathbf{Y_{mis(t=2)}}|\mathbf{Y_{obs}}$.

The P-step: After missing data have been imputed for each subject in the I-step, the full data (both observed and the newly imputed values) are used to estimate the posterior probability means and covariances. The posterior is estimated by Bayes' formula. These are the new parameters used in the subsequent I-step. The prior can be either an informative or uninformative prior.

Thus, these two steps form a chain where at each iteration, q, the parameter estimates, $\theta^{(q)}$ are used to generate new imputes, $\mathbf{Y^{(q+1)}_{mis(t=1)}}, \mathbf{Y^{(q+1)}_{mis(t=2)}}$, from the conditional distribution, $\mathbf{Y_{mis(t=1)}}, \mathbf{Y_{mis(t=2)}}|\mathbf{Y^{(q)}_{obs}}$. The subsequent P-step estimates $\theta^{(q+1)}$ from the posterior $p(\theta|\mathbf{Y_{obs}}, \mathbf{Y_{mis(t=1)}}^{(q+1)}, \mathbf{Y_{mis(t=2)}}^{(q+1)})$. The steps are continued until the missing value imputations approximate independent draws from $\mathbf{Y_{mis}}|\mathbf{Y_{obs}}$, and the desired number of multiply imputed complete data sets is reached.

If no prior information is available, as is likely true for an experimental treatment, a noninformative Jeffrey's prior may be used. This generates a covariance matrix distributed as an inverse Wishart, $\Sigma|Y \sim W^{-1}(n-1, (n-1)S)$, where S is the sample variance-covariance matrix, n is the number of subjects in the treatment arm, and multivariate normal mean vector $\mu|(\Sigma, Y) \sim N(\bar{y}, \frac{1}{n}\Sigma)$.

Given that in non-inferiority trials, the treatment used in the active control arm has necessarily been studied in previous trials, historic data can be utilized to improve confidence in imputed values. This approach makes several strong assumptions. That estimands targeted by historic trials are the same as those in the current trial, and that furthermore, the predominant missing mechanism is correctly accounted for in historic trials and that the mechanism remain constant across trials. In practice, it may be more conservative to use a prior with a higher than expected treatment effect in the active control to steepen the burden of proof.

With available historic information on both the covariance and the mean effect of the active control, the analyst can set priors on the parameters in Σ and μ via the hyperparameters S_0 and μ_0 respectively. The posterior distributions are then estimated from

$$\begin{aligned}\Sigma|Y \sim & W^{-1}(n+n_0-1,(n-1)S+(n_0-1)S_0+S_m),\\ \mu|(\Sigma,Y) \sim & N(\frac{1}{n+n_0}(n\bar{y}+n_0\mu_0),\frac{1}{n+n_0}\Sigma),\end{aligned}$$

where n is the sample size within the treatment arm, n_0 is a hyperparameter representing the number of subjects used for prior parameter estimates, and $S_m = \frac{nn_0}{n+n_0}(\bar{y}-\mu_0)(\bar{y}-\mu_0)'$.

5.2.5 Tipping-Point Analysis

A tipping point analysis is a common procedure for evaluating the sensitivity of conclusions to missing data assumptions[37]. The procedure involves repeated analyses over a range of some parameter in the model to observe any change in conclusions. A common implementation of a tipping point analysis is to perform multiple imputation under an MAR assumption and adjust the imputed values by a scale or shift parameter. Sometimes, the objective of a tipping-point analysis is to seek a tipping point within the range of values whereupon the study conclusions change. Trialists then consider the relative likelihood of the scenario represented by the tipping point.

The pattern-mixture models introduced in the previous section can serve as a starting point for a tipping point analysis. The identifying restrictions then become

$$\mu^{(2)}_{ET,2} = \mu^{(1)}_{ET,2} = \mu^{(0)}_{ET,2} + c_{ET}\delta/2 \tag{5.4}$$

$$\mu^{(2)}_{AC,2} = \mu^{(1)}_{AC,2} = \mu^{(0)}_{AC,2} + c_{AC}\delta/2 \tag{5.5}$$

where $c_s \in (-1, 1)$ and we allow the c_s's to vary independently. This approach allows us to investigate multiple scenarios ranging in magnitude along two dimensions: departure from the MAR assumptions and degree of treatment arm interaction with the MNAR mechanism. These two considerations are helpful in evaluating the plausibility of outcomes. Deviation from MAR at the final timepoint (for a given treatment arm) is captured by $|c_s|$, with maximal deviation when $|c_s| = 1$. Treatment arm interaction is represented by the absolute difference, $|c_{ET} - c_{AC}|$, with larger values indicative of a stronger interaction effect. We demonstrate a tipping point approach on simulated data in Section 5.3.4.

5.3 Simulations

We investigate the performance of a sensitivity analysis using several versions of our proposed pattern-mixture model approach developed in the previous section. Using simulations, we aim to show these approaches can assist researchers in determining departures from the MAR assumption and their effect on non-inferiority trial conclusions. Specifically, a robust sensitivity analysis will:

a) under the null hypothesis, identify MNAR scenarios that call into question an incorrect conclusion of non-inferiority in the primary analysis, and

b) under the alternative, if the true difference in means is equal to zero (i.e. the treatment is non-inferior), then the sensitivity analysis is likely to support the primary analysis.

While the methods we proposed are expected to produce conservative results, we aim to show the degree of their conservatism and that evaluated in tandem with a tipping point

analysis, the models provide valuable insight into the plausibility of alternative, not-at-random missing data scenarios.

Against the results from a primary analysis using the MMRM, we compared three versions of the pattern-mixture models fit via multiple imputation: (1) multiple imputation with worst-case shift, (2) multiple imputation with an informative prior on the means of the active control, and (3) multiple imputation with an informative prior on the active control and worst-case shift in the experimental treatment. These approaches are described in detail below.

In Section 5.3.4 we demonstrate the approaches in an enhanced tipping-point analysis with practical visualization of results.

5.3.1 Data Generation

Outcome data were generated from a multivariate normal distribution with a non-linear, increasing trajectory of means. The trial design consisted of measurements taken at baseline and at four subsequent times. For each subject i, the vector of outcomes $\mathbf{Y_i}$ was drawn from $MNV(\boldsymbol{X_i\beta_h}, \boldsymbol{\Sigma})$ with $\boldsymbol{X_i}$ the 5 x 10 design matrix, $\boldsymbol{\beta_h}$ a 10 x 1 vector corresponding to the means at each timepoint for both treatment arms, and $\boldsymbol{\Sigma}$ the 5 x 5 variance-covariance matrix with heterogeneous autoregressive structure.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 & \sigma_1\sigma_5\rho^4 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 & \sigma_2\sigma_5\rho^3 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho^2 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho \\ & & & & \sigma_5^2 \end{pmatrix}$$

with $\sigma_1 = 1.5$,$\sigma_2 = 1.75$,$\sigma_3 = 2.0$, $\sigma_4 = 2.25$, $\sigma_4 = 2.5$, and $\rho = 0.7$. Data were generated under both the null hypothesis, where treatment arm means differed by 1 at the final timepoint in $\boldsymbol{\beta_{h_0} = (\mu_{ET(h_0)}, \mu_{AC})}$, and the alternative hypothesis of non-inferiority, where

the treatment arm means were the same at every time in $\boldsymbol{\beta_{h_a} = (\mu_{ET(h_a)}, \mu_{AC})}$:

$$\text{(Under } H_0\text{)} : \boldsymbol{\mu_{ET(h_0)}} = (0\,0.32\,0.40\,0.45\,0.5),\ \boldsymbol{\mu_{AC}} = (0\,0.95\,1.19\,1.36\,1.5)$$

$$\text{(Under } H_a\text{)} : \boldsymbol{\mu_{ET(h_a)}} = (0\,0.95\,1.19\,1.36\,1.5),\ \boldsymbol{\mu_{AC}} = (0\,0.95\,1.19\,1.36\,1.5)$$

The sample size used was 216, or 108 per treatment arm to achieve 90% power to show non-inferiority with a pre-specified non-inferiority margin,[34].

Missing Data

After generating the complete data, we simulated MAR and MNAR mechanisms by removing values at times 2, 3 and 4. For MAR, data we flagged measurements as missing with probability proportional to the z-score of the previous observation. Z-scores were computed with data from both treatment arms combined. Data MNAR were flagged in a similar manner except that probabilities were based on current values.

In addition, we modeled each mechanism both with and without a treatment arm interaction. Without interaction, low values were more likely to be flagged missing. With interaction, low values in the experimental treatment arm were more likely missing, and high values in the active control were more likely missing. All measurements proceeding the flagged values were also flagged as missing to induce a monotone pattern. We defined the proportion of missing data as the number of missing observations at the final time point divided by the total number of subjects. Tuning parameters were used to achieve the correct proportion of missing data, either 20 or 40 percent. These proportions were selected to be in agreement with real longitudinal clinical trials[14,17].

5.3.2 Analysis, Performance Measures and Visualization

The primary goal of the simulation study was to compare conclusions from the primary analysis with sensitivity analyses implemented via multiple imputation (MI). For direct comparison with the primary MMRM, we analyzed results from the following three imputation models:

Method 1: MI by regression method and worst-case shift. This approach imputes monotone missing data using the regression method then shifts the imputed values at the final time in the active control by $\delta/2$ and in the experimental treatment by $-\delta/2$.

Method 2: MI with informative prior. This model imputes missing data in the active control arm with the MCMC method, placing an informative prior on the means based on historic data. In our simulations, the hyperparameters of the informative prior were the means and the variance-covariance matrix used in the data-generation model, i.e. we assumed $\boldsymbol{\beta_{h(Activecontrol)}} \sim MVN(\boldsymbol{\mu_{AC}}, \boldsymbol{\Sigma})$. The prior sample size was taken to be the same size as the current study. The experimental treatment arm is imputed using regression MI.

Method 3: MI with informative prior and shift in experimental treatment arm. This method combines two approaches above. In the active control, missing data are imputed in the same manner as in method 2 with MCMC MI. Values in the experimental treatment arm are imputed using regression MI and they are shifted by $-\delta/2$.

In each case, we generated 100 imputations, the complete data sets were analyzed by MMRM and the difference in change from baseline at the final timepoint was estimated between arms. Estimates and standard errors were then combined according to Rubins rules[35]. We concluded non-inferiority if the upper limit of the 90% confidence interval was less than 1. With no missing data, this test of non-inferiority controls type I error at 5%.

We generated 500 replicates of each missing data mechanism: MAR, MAR with treatment arm interaction, MNAR, and MNAR with treatment arm interaction. The missing data proportion was fixed at 20% under both the null and alternative hypothesis.

For comparison of methods 1-3 with the primary analysis, we calculated bias, rejection % (type I error or power) and coverage of the true effect difference under the 90% confidence interval for both the null and alternative hypotheses. We also estimated the empirical standard error, the variance of estimates across replicates, and Monte Carlo standard errors to quantify simulation uncertainty due to using a finite number of replicates[49].

In addition to simulations of methods 1-3, we conducted an enhanced tipping point analysis by extending methods 1 and 3. For method 1, we reanalyzed each simulated data

set with shift parameters to varying in increments of 0.2 from $-1/2$ to $1/2$ independently by treatment arm. For method 3, we modeled the missing data that same with an informative prior (i.e. the active control arm was not shifted), but varied the shift parameter for the missing experimental treatment data from $-1/2$ to $1/2$. These results are reported in Section 5.3.4.

5.3.3 Results

Tables 5.3.3 and 5.3.3 compare the average performance for the primary analysis with MMRM and sensitivity analysis methods 1, 2, and 3 using 500 replicates for each missing data scenario with 20 percent missing. In the primary analysis, results were unbiased except when the missing mechanism was MNAR with arm interaction. This is an expected result as the MMRM is known to produce unbiased estimates when data are MAR. Even when the data are MNAR, the MMRM performs well. This is likely because the missing mechanism we used was not strictly MNAR. Although the probability of missingness was determined only from the missing values themselves, the high correlation between repeated measures meant that a MAR mechanism could explain the missingness at least in part.

Table 5.2: Estimates of bias for the primary analysis with MMRM and three sensitivity analysis models under the null and alternative hypothesis and four missing data mechanisms. Method 1: worst case shift; method 2: informative prior for AC; method 3: informative prior for AC and $-\delta/2$ shift in ET. Negative bias indicates underestimation of the treatment difference $\Delta = (\mu_{AC,4} - \mu_{AC,0}) - (\mu_{ET,4} - \mu_{ET,0})$.

Scenario $H_0 : \Delta = 1$	MAR	MAR with interaction	MNAR	MNAR with interaction
Primary Analysis	-0.036	0.007	-0.007	-0.369
Method 1	0.160	0.202	0.186	-0.172
Method 2	-0.037	0.040	-0.015	-0.324
Method 3	0.062	0.141	0.081	-0.227
Scenario $H_a : \Delta = 0$	MAR	MAR with interaction	MNAR	MNAR with interaction
Primary Analysis	-0.012	0.016	0.010	-0.361
Method 1	0.185	0.209	0.205	-0.166
Method 2	-0.011	0.044	-0.000	-0.318
Method 3	0.089	0.141	0.097	-0.221

Compared to the primary analysis, methods 1-3 were biased in the direction of the null, that is, they overestimated (or underestimated less) the difference in treatment effects. Methods 1 and 3 were over-conservative with estimated type I errors around 3% (Table 5.3.3). However, under the alternative, method 1 produces the lowest estimated power under MAR with treatment arm interaction at 64%. Under the most extreme missing

mechanism, MNAR with interaction, none of the sensitivity analysis models could control type I error or produce estimates that were unbiased. Method 2 with informative prior on the active control did not produce different results from the primary analysis. Coverage values were over 80% except under the extreme case, MNAR with treatment arm interaction.

Empirical standard errors and model standard errors were similar at approximately 0.40, indicating that our simulations were behaving reasonably. Monte Carlo standard errors ranged from 0.007 (7%) to 0.02 (2%).

Table 5.3: Power, type I error and coverage for the primary analysis with MMRM and 3 sensitivity analysis models under the null and alternative hypothesis. Method 1: worst case shift; method 2: informative prior for AC; method 3: informative prior for AC and $-\delta/2$ shift in ET; T1E: type 1 error. Expected values are: type I error=5%; coverage=90%; power=90%. All table values are percentages.

Scenario $H_0 : \Delta = 1$	MAR		MAR with interaction		MNAR		MNAR with interaction	
	T1E	Coverage	T1E	Coverage	T1E	Coverage	T1E	Coverage
Primary Analysis	5.4	91.8	4.0	90.2	6.4	87.2	25.6	73.8
Method 1	2.0	88.0	1.4	86.2	2.6	85.8	13.6	85.2
Method 2	6.5	90.6	3.6	89.4	6.6	86.6	23.4	75.6
Method 3	3.2	90.6	1.8	87.8	4.6	86.2	17.6	81.2
Scenario $H_1 : \Delta = 0$	MAR		MAR with interaction		MNAR		MNAR with interaction	
	Power	Coverage	Power	Coverage	Power	Coverage	Power	Coverage
Primary Analysis	82.4	88.8	81.8	89.6	81.2	88.2	97.6	74.8
Method 1	70.0	84.4	63.6	85.4	67.6	82.6	91.8	87.2
Method 2	85.0	87.2	80.0	88.8	83.4	88.4	97.2	76.8
Method 3	78.0	87.0	71.2	87.2	76.4	86.2	94.4	82.0

5.3.4 Demonstration of Tipping Point Analysis

We demonstrate an application of our model as an enhanced tipping point analysis with graphical display of the results. Liublinska and Rubin[29] developed the approach to aide researchers in visualizing conclusions under a range of plausible outcomes for missing binary data in a two-armed RCT. We adopt the technique to display results from two approaches to sensitivity analysis for simulated data:

1. Extending method 1, every simulated data set was reanalyzed by multiple imputation for 121 models: the mean shifts in imputed values varied in increments of 0.2 from $-1/2$ to $1/2$ independently by treatment arm.

2. Extending method 3, every simulated data set was reanalyzed by multiple imputation for 11 models. The active control imputes were drawn from the posterior predictive

distribution given the same informative prior (described above), and the imputes on the experimental treatment arm were systematically shifted from $-1/2$ to $1/2$ in increments of 0.2.

Figures 5.1 and 5.2 demonstrate the results from sensitivity analyses of selected simulated data sets when there is 40 percent missing data under four mechanisms: MAR, MAR with treatment arm interaction, MNAR, and MNAR with treatment arm interaction. Figure 5.1 demonstrated selected simulated datasets under the null hypothesis (with $\Delta = 1$) and Figure 5.2 datasets are drawn from the alternative (with $\Delta = 0$). Each grid displays p-values from the 121 analyses on a single data set for every combination of shift by treatment arm. Figures 5.3 and 5.4 show results for the analogous tipping point analysis with a shift only in the experimental treatment arm; the control arm is always imputed given the informative prior using the MCMC method for multiple imputation.

Statisticians can examine these displays with clinicians or other subject experts and decide whether they represent realistic scenarios. It is easy to conclude, in some case, that even with extreme assumptions about distributional differences between the missing and non-missing cases, the estimates obtained from the sensitivity analysis are consistent with the primary analysis. Of course, in our demonstrations, the simulated missing mechanisms are likely more extreme than in practice. Furthermore, our examples contained 40% which may be extreme in an RCT when the experimental treatment is a drug, device or procedure. Follow-up times tend to be shorter and fewer subjects drop out in these trials than in behavioral trials for instance[50].

One advantage to this visualization is that cases that are ambiguous or close to the boundary of the alternative will tend to reveal this ambiguity in sensitivity analysis. Even the results of an unclear sensitivity analysis can be reported alongside a primary analysis to suggest caution in generalizing the findings. If researchers suspect a MNAR mechanism, one could make an argument for requiring all the tipping point analyses be significant for a conclusion of non-inferiority. This is certainly a conservative approach and it may be appropriate for some applications.

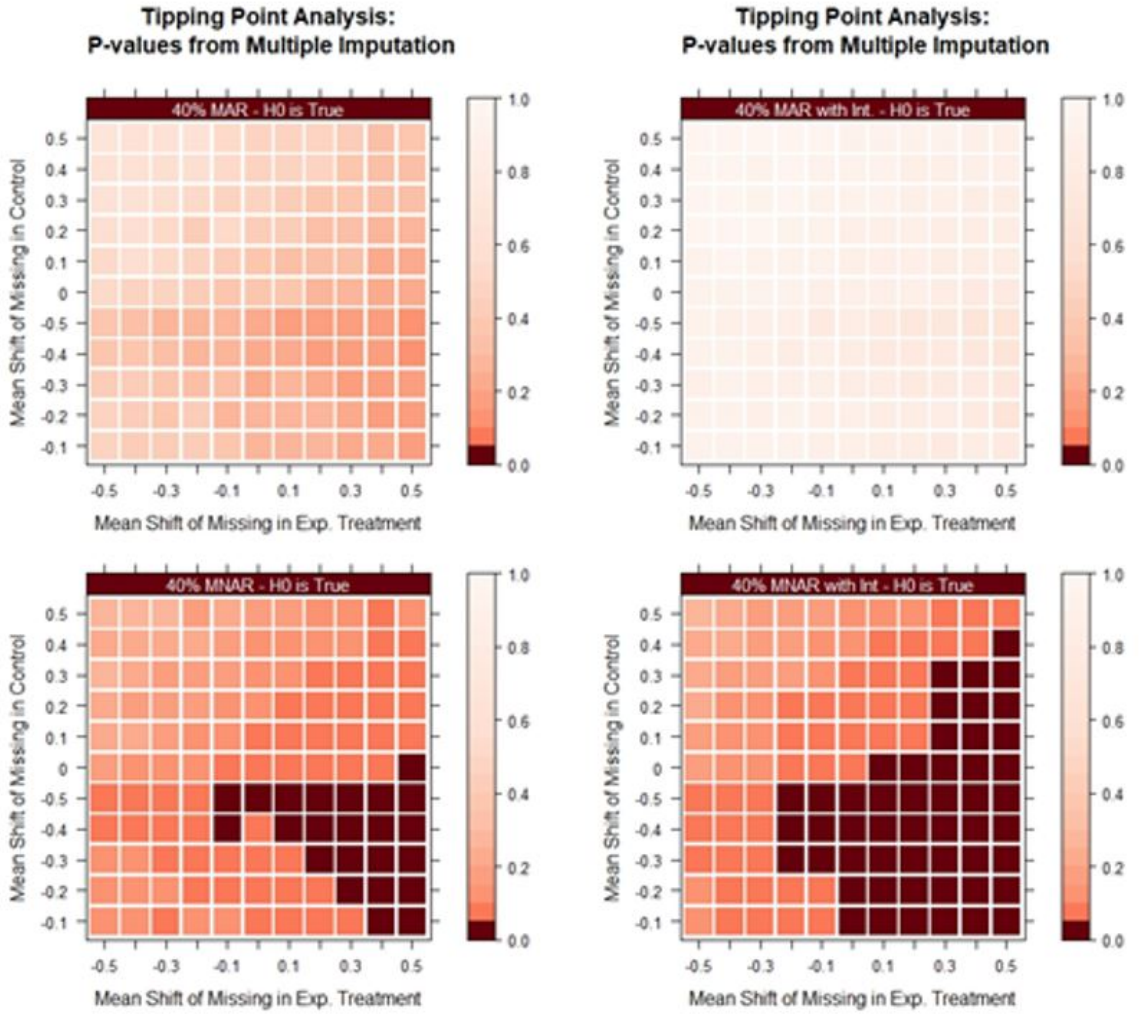

Figure 5.1: P-values from a tipping point analysis when H_0 is true. Each of the four grids is a sensitivity analysis on a single, simulated data set under the null hypothesis. The labels on the x-axis indicate the mean shift of imputed values for the experimental treatment. Labels on the y-axis are the mean shift of imputed values for the active control. The p-values are calculated from a multiple imputation analysis with a regression imputation model and the MMRM as primary analysis. The colorkey assigns a darker value to greater evidence against the null hypothesis, so that the darkest red squares in the grid are statistically significant ($p < 0.05$). Greater numbers of dark squares support a conclusion of non-inferiority. Top-left: 40% MAR. Top-right: 40% MAR with arm interaction. Bottom-left: 40% MNAR. Bottom-right: 40% MNAR with arm interaction.

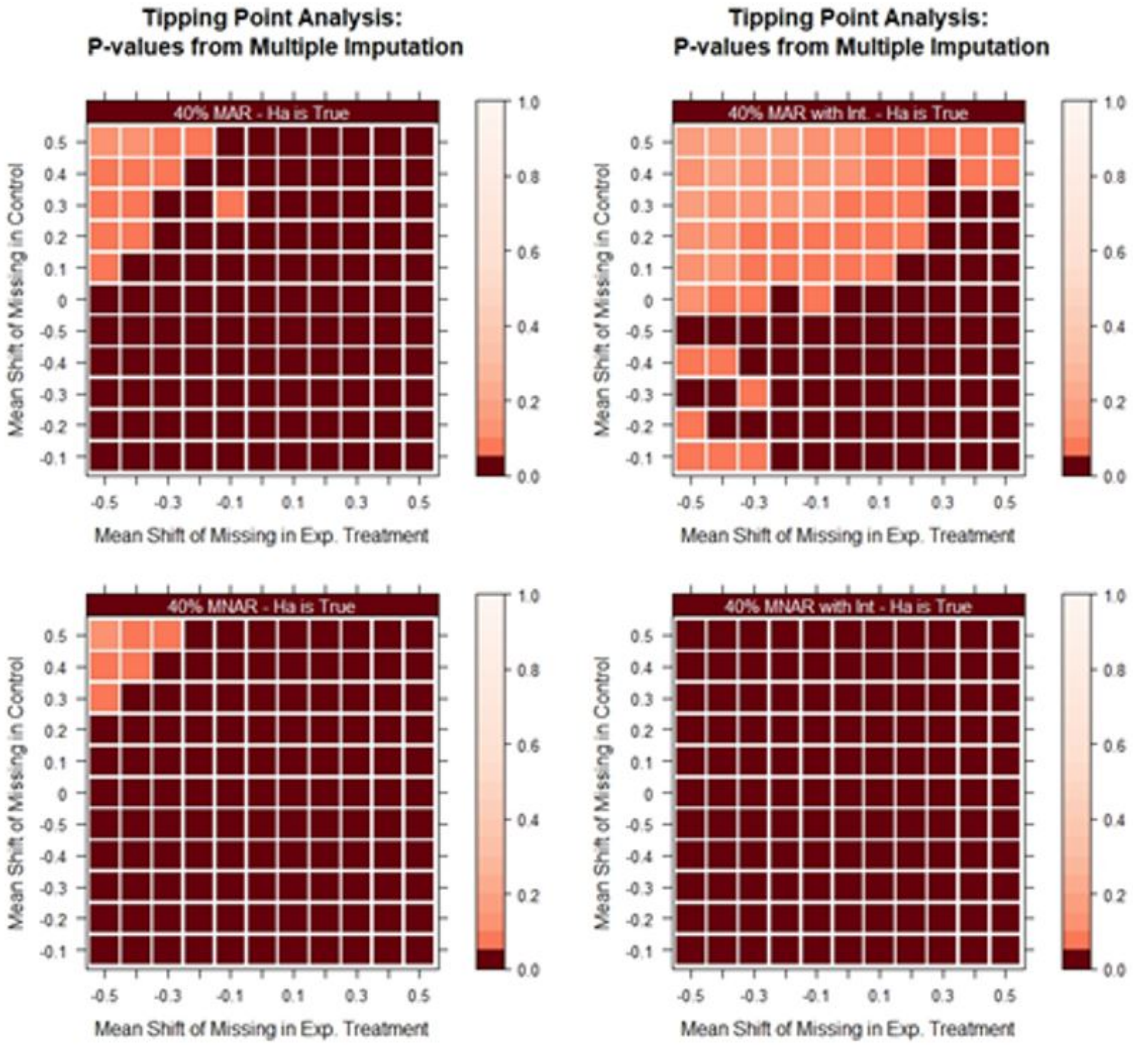

Figure 5.2: P-values from a tipping point analysis when H_a is true. Each of the four grids is a sensitivity analysis on a single, simulated data set under the alternative hypothesis where the true difference in effects is zero. The labels on the x-axis indicate the mean shift of imputed values for the experimental treatment. Labels on the y-axis are the mean shift of imputed values for the active control. The p-values are calculated from a multiple imputation analysis with a regression imputation model and the MMRM as primary analysis. The colorkey assigns a darker value to greater evidence against the null hypothesis, so that the darkest red squares in the grid are statistically significant ($p < 0.05$). Greater numbers of dark squares support a conclusion of non-inferiority. Top-left: 40% MAR. Top-right: 40% MAR with arm interaction. Bottom-left: 40% MNAR. Bottom-right: 40% MNAR with arm interaction.

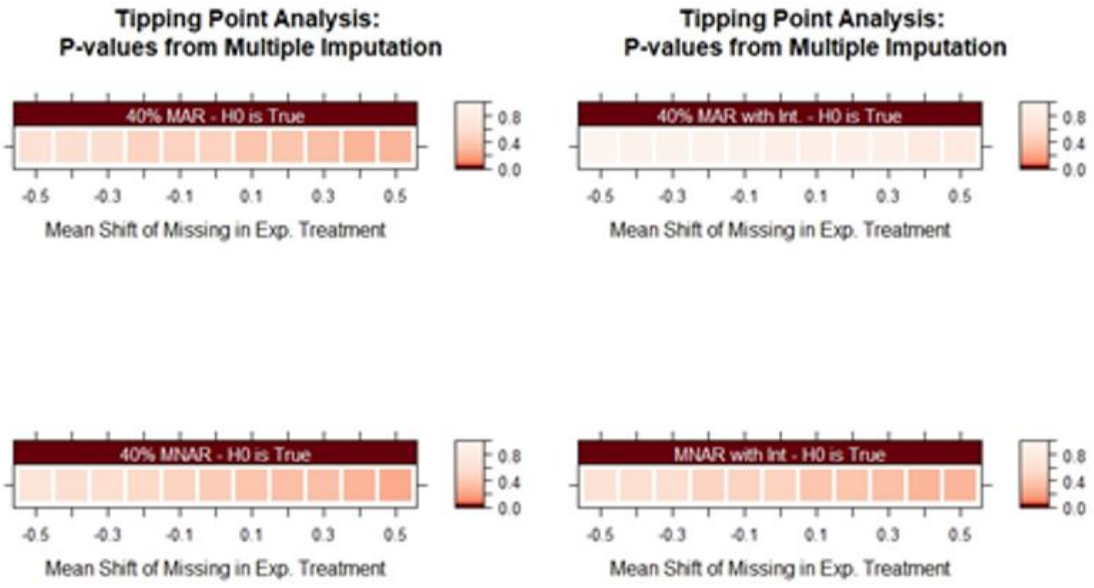

Figure 5.3: P-values from a tipping point analysis when H_0 is true. Each of the four grids is a sensitivity analysis on a single, simulated data set under the null hypothesis. The p-values are calculated from a multiple imputation analysis that imputes separately by treatment arm. The active control imputes are drawn from the posterior predictive distribution based on an informative prior, and the experimental treatment imputes are drawn from a regression imputation method. The analysis model is the MMRM. The labels along the x-axis indicate the mean shift of imputed values for the experimental treatment arm. The colorkey assigns a darker value to greater evidence against the null hypothesis, so that the darkest red squares in the grid are statistically significant ($p < 0.05$). Greater numbers of dark squares support a conclusion of non-inferiority. Top-left: 40% MAR. Top-right: 40% MAR with arm interaction. Bottom-left: 40% MNAR. Bottom-right: 40% MNAR with arm interaction.

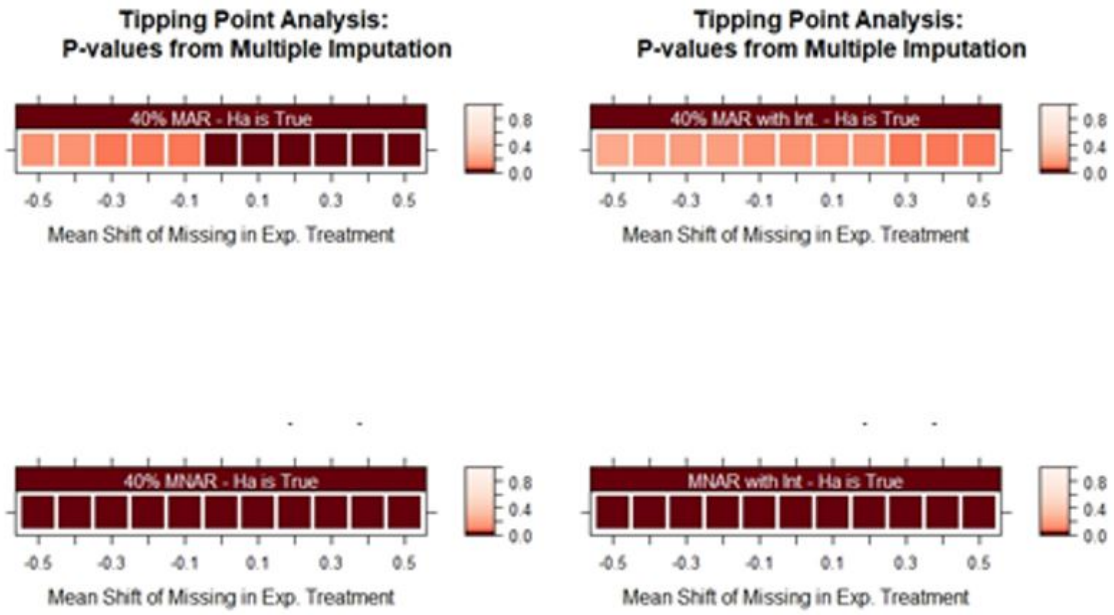

Figure 5.4: P-values from a tipping point analysis when H_a is true. Each of the four grids is a sensitivity analysis on a single, simulated data set when both treatment arms follow the same trajectory of means. The p-values are calculated from a multiple imputation analysis that imputes separately by treatment arm. The active control imputes are drawn from the posterior predictive distribution based on an informative prior, and the experimental treatment imputes are drawn from a regression imputation method. The analysis model is the MMRM. The labels along the x-axis indicate the mean shift of imputed values for the experimental treatment arm. The colorkey assigns a darker value to greater evidence against the null hypothesis, so that the darkest red squares in the grid are statistically significant ($p < 0.05$). Greater numbers of dark squares support a conclusion of non-inferiority. Top-left: 40% MAR. Top-right: 40% MAR with arm interaction. Bottom-left: 40% MNAR. Bottom-right: 40% MNAR with arm interaction.

5.4 Discussion

We have developed a pattern-mixture model approach and tipping point analysis for sensitivity analysis with respect to missing data assumptions in a non-inferiority trial. We have shown its application as an imputation model within a multiple imputation framework so that the analysis model can be consistent with the primary analysis. We have shown how the tipping point analysis can be conceived as investigating extremes in the missing mechanism departure from MAR and treatment arm interaction. Furthermore, we have presented the approach using a strong, informative prior on the means of the treatment effect in the active control arm based on expert opinion or historical data. Adapting the visualizations of Kim et al.[30], we provided a graphical tool for assessing the sensitivity of conclusions to plausible assumptions about the distribution of missing data.

We showed that with minimal loss of power, the regression multiple imputation with worst-case shift gives conservative estimates and better controls type I error when compared with a MMRM primary analysis when data are MNAR and the missing mechanism has treatment arm interaction. The MCMC multiple imputation method with an informative prior on the active control can be used in combination with shift parameters in the experimental treatment arm and combined for a conservative decision rule, provided selection of prior information is based on a reasonable assumption of constancy between the historic and present clinical trial.

Using the pattern-mixture formulation with an imputation model is preferable to direct estimation of treatment effects with a model that assumes a MNAR mechanism and is vulnerable to misspecification. This is of concern in MNAR modeling when strong distributional assumptions are made (as the observed data are assumed not to carry information about the missing values).

The National Research Council[15] report on missing data in clinical trials emphasizes the importance of avoiding missing data altogether insofar as trialists are able. Since then, there have indeed been growing awareness of the importance to minimize patient dropout and improvements in the collection of outcome measures even upon treatment cessation, adverse

events, or other protocol violations. Nevertheless, missing data are nearly unavoidable and should be planned for in the design stage[14,17].

How analysts handle the missing data is key, and ideally plans have been laid out from the design stage including pre-specification of how missing data will be imputed, whether subjects who have discontinued treatment, or had adverse events will be included in the analysis. In this study, we have assumed that missing data are truly missing, that is, measurements have not been collected; although, the sensitivity analysis we propose may also be used to examine counterfactual estimands (where the model imputes as though adverse events and non-compliance had not occurred). Such estimands may indeed be conservative in the non-inferiority setting, particularly in the case of non-compliance (see Chapter 4).

Some tipping-point analyses in the literature are so-named because the objective is to literally search for a parametric value that reverses the study conclusions[29]. We argue that this approach is less applicable or versatile in NI trials where the assumptions about the differences in treatment arms can have big consequences for the conservatism or anti-conservatism of conclusions. Searching for tipping points in a two-dimensional space is more complicated, hence, our approach begins with the most plausible assumption under the null hypothesis, where treatment effects are expected to differ by the non-inferiority margin and varies the shift parameters within that extreme case.

If the imputation model is estimated from an informative prior on the active control, the prior parameters should be selected thoughtfully and informed by results of historic trials. Eliciting priors is recommended with help from expert advice[45]. In addition, non-inferiority trials should have a wealth of data showing the control as an effective treatment. Indeed, without the assumption that the active control performs equally well, a non-inferiority trial risks losing assay sensitivity[5], and a conclusion of non-inferiority would not imply efficacy of the experimental treatment. Modeling the missing data in the active control arm as though subjects responded as expected based on historic trials helps preserve the assumption of assay sensitivity and is more in line with a conservative, intention-to-treat analysis.

We limited our study to a few simple missing data mechanisms, and we did not model any covariate-dependent dropout. In practice, incomplete data are likely to result from

a complex mechanism that involves multiple variables, some of which may be unobserved. The lessons from much of the research on missing data emphasize that the more information collected the better. These auxiliary data can then be used to determine, first, whether a subject should be included in the analysis and, next, how to align assumptions about missing data with the primary estimand. As trialists improve data collection practices, analysts can use increasingly sophisticated statistical techniques to understand relationships between missingness and auxiliary variables. As these relationships are better understood, analysts are equipped to deal with missing data better and build detailed models of the missing mechanism predominantly at play.

In this study, we only evaluated the case of continuous outcomes. Other authors have proposed tipping-point procedures for binary outcomes[51] and similar visualizations can be produced for counterfactual estimands in non-inferiority trials[30]. Also, our pattern-mixture model only accounts for MNAR data at the final timepoint. This may not be reasonable for studies with many data collection periods and with dropout early in the trial. Further research is needed to generalize the pattern-mixture model for informative dropout at any timepoint.

Early literature on the statistics of non-inferiority trials has examined the selection of analysis sets[12,21,26]. Simulations showed that intention-to-treat (ITT) analysis could inflate type I error rates in the presence of missing data and treatment non-compliance. On the other hand, per-protocol analyses that exclude subjects with missing observations or protocol violations do not reliably control type I error and are subject to selection bias. More recently, guidelines from regulatory bodies have de-emphasized the focus on ITT and shifted to specification of estimands[2,6,52]. An estimand, that which is to be estimated, is typically an unknown parameter of scientific interest defined via the target population (i.e. analysis set), the measure of outcomes, the handling of post-randomization events, and the analysis method. This shift underscores the need for analysts to handle missing data in a manner consistent with the trial estimand and to explore the impact of missing data assumptions upon conclusions. However, to date, there is a paucity of research concerning sensitivity analysis with respect to missing data assumptions in non-inferiority trials.

5.5 Recommendations

For longitudinal, continuous outcomes in non-inferiority trials, we recommend the use of a MMRM as the primary analysis because it produces unbiased estimates for data MAR and is not vulnerable to misspecification. We do not advocate the use of a MNAR pattern-mixture model for primary analysis. Since data MAR are unlikely to bias the estimate of effect and, in reality, missing data are explained by a combination of mechanisms, the assumptions implicit in our proposed pattern-mixture model of pure not at random missingness are likely too extreme for a primary analysis. As part of a sensitivity analysis, defense of the conclusions based on a MNAR model should be made on clinical, not statistical grounds[40]. The regression and MCMC multiple imputation methods proposed are complementary approaches that can be employed in sensitivity analysis with respect to assumptions about missing data and protocol violations in the primary analysis. We emphasize that the real utility of the proposed approaches is how they enable researchers to inspect their conclusions under various plausible missing data assumptions in a way that is convenient and easy to interpret.

Chapter 6

Conclusions and Future Work

In this dissertation, I have demonstrated the relevance of missing data problems in non-inferiority trials both in theory and practice. My systematic review examined missing data handling by biostatisticians contributing to the body of biomedical research. I investigated the choice of analysis set on non-inferiority conclusions when subjects are non-compliant or withdraw from the trial. Methods proposed for handling missing data were compared using Monte Carlo simulations and demonstrated with examples. I proposed several models under appropriate assumptions for the non-inferiority setting which on average yielded conservative estimates of effect and control of type I error.

This work supports the thesis that methods for missing data handling must be considered in the context of trial design. Methods that may be appropriate under some trial conditions will inflate type I error under other conditions. Further, the assumptions an analyst makes about the missing data mechanism, whether implicit or explicit, will have consequences for the interpretation and reproducibility of results.

6.1 Missing Data Handling in Practice

With my systematic review of non-inferiority trials, I found evidence that missing data remain problematic in trials and that statistically unprincipled methods for handling them are the norm. While these results should not be celebrated, the growing number of publications devoted to statistical methods for non-inferiority design should at least give us some

reassurance. As methodologists publish their research, the industries and regulatory bodies that depend on these trials will become increasingly aware of the issues.

The systematic review stands out among other reviews as one that looks specifically at missing data for non-inferiority trials. While missing data are indeed problematic in all data analysis, many statistical practices have been adopted for their believed conservatism in the superiority setting. Hence, it was essential for statisticians to assess whether researchers were aware of the non-inferiority design's susceptibility to false positives.

It is natural to wonder whether practices will improve in the analysis of non-inferiority designs. Since the trials in my review were published, the ICH has published further on estimands and the FDA completed its final guidance document on analysis and design of non-inferiority trials. Furthermore, there has been a shift away from the nomenclature of analysis sets toward the concept of estimands. This should improve practice generally because it encourages pre-specification of analysis and consideration of how missing data and other protocol deviations should be handled.

6.2 In Search of Meaningful Estimands

The discovery that many trialists reported both the ITT and per-protocol analysis sets led to the basis for my investigation of multiple imputation methods. It is still unclear whether recommendations for any analysis can be universally applied. One hopes that analyzing the data as randomized will produce conservative estimates upon which doctors and patients can make sound decisions; however, some patients will want to know how they will fare if they are among those who take medication as directed. An ITT analysis may not target this estimate, and as our simulations showed, may lead to greater confidence in the treatment than is warranted.

The control-based imputation approach I implemented targets a hypothetical estimand that imagines a fully compliant sample. This approach was more conservative and was unbiased for the treatment effect in the compliant population. The implication is that treatment decisions may well depend on more than simply the ideal biological efficacy as

estimated by a perfectly compliant sample. Treatment decisions may be optimized by conditioning on the expected compliance of an individual.

6.3 Sensitivity Analysis

In the final chapter of this work, I presented a model for investigating the sensitivity of conclusions to assumptions about missing data: one that is intuitive and agrees with the recommended approach to impute under the assumptions of the null hypothesis; although, possibly not that extreme. As in Kim et al.[30], the same approach can be used for sensitivity analysis with respect to the analysis set if the data are compromised by non-adherence or other protocol deviations.

While statistical methods have room for improvement, they are limited by what sources of error are unknown and cannot realistically be accounted for in analysis. Robust trial conclusions result from efforts on multiple fronts to improve trial conduct, including patient retention and continued data collection even upon discontinuation of treatment. Statisticians can then decide on the appropriate analysis given the scientific questions and be unencumbered by the need to assume a missing data mechanism at all.

6.4 Future Research

Future research may be motivated by the limitations of the existing body of work; however it should be relevant to the needs of patients, clinicians and stakeholders who rely on products developed using the non-inferiority trial design. Given the increased usage of non-inferiority designs and academic interest in their statistical challenges, we would expect practice to have evolved over the last several years. Therefore, a follow-up systematic review would complement the earlier review and allow comparisons of relevant statistics on missing data and how they are handled. I hypothesize that reporting of assumptions and the analysis methods will on average be improved.

To draw meaningful conclusions from non-inferiority trials, logic dictates we must rely on the assumptions of constancy and assay sensitivity. Assay sensitivity is the trial's ability to distinguish an effective treatment from a placebo, and constancy requires that conditions

in the non-inferiority trial are sufficiently close to those of the historical trials so that the full treatment effect of the active control is maintained. While missing data may reduce power and produce biased estimates in any trial design, in non-inferiority trials, missing data may weaken assumptions of constancy and assay sensitivity. To my knowledge, there is no published literature directly investigating the effect of missing data handling on these assumptions. New research could aim to determine, via Monte Carlo simulations, whether methods for handling missing data in non-inferiority trials undermine constancy and assay sensitivity.

There is a need for further research into Bayesian approaches to non-inferiority trial design and analysis. Bayesian methods provide a natural way to incorporate historical data in the form of prior distributions. Gamalo et al.[27] presented a Bayesian approach for determining evidence of drug effect (to determine a margin), testing the hypothesis, and estimating sample size. The sensitivity analysis method presented in Chapter 5 demonstrated a Bayesian approach to missing data imputation, and this complements a frequentist approach. However, a Bayesian paradigm might be more readily accepted by regulatory bodies after research investigates the frequentist operating characteristics of the Bayesian analysis.

Finally, the pattern-mixture model presented in Chapter 5 explores a MNAR mechanism that occurs at the final timepoint. In reality, MNAR data may occur at any time, and indeed, are likely to reflect occurrences at the time of dropout, not only with respect to the end of the trial period. Hence, a sophisticated extension of this approach might introduce weights for the shift parameter in proportion to the time of dropout, but the details of a robust sensitivity model will depend on the circumstances of the individual trial.

Bibliography

[1] Mark D Rothmann, Brian L Wiens, and Ivan SF Chan. *Design and analysis of non-inferiority trials*. Chapman and Hall/CRC, 2016.

[2] U.S. Food and Drug Administration. Non-inferiority clinical trials to establish effectiveness: Guidance for industry. `https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf`, 2016.

[3] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[4] Andrew D Garrett. Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine*, 22(5):741–762, 2003.

[5] Brian L. Wiens and Gerd K. Rosenkranz. Missing data in noninferiority trials. *Statistics in Biopharmaceutical Research*, 5(4):383–393, nov 2013. doi: 10.1080/19466315.2013.847383. URL `http://dx.doi.org/10.1080/19466315.2013.847383`.

[6] International Conference on Harmonisation. E9 (R1) Estimands and Sensitivity Analysis in Clinical Trials. *Guidance*, 9(June):9, 2017.

[7] Grace Wangge, Olaf H. Klungel, Kit C.B. Roes, Anthonius de Boer, Arno W. Hoes, and Mirjam J. Knol. Should non-inferiority drug trials be banned altogether? *Drug Discovery Today*, 18(11-12):601–604, jun 2013. doi: 10.1016/j.drudis.2013.01.003. URL `http://dx.doi.org/10.1016/j.drudis.2013.01.003`.

[8] Grace Wangge, Olaf H. Klungel, Kit C. B. Roes, Anthonius de Boer, Arno W. Hoes, and Mirjam J. Knol. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: A systematic review. *PLoS ONE*, 5(10):e13550, oct 2010. doi: 10.1371/journal.pone.0013550. URL `http://dx.doi.org/10.1371/journal.pone.0013550`.

[9] Mouna Akacha, Frank Bretz, David Ohlssen, Gerd Rosenkranz, and Heinz Schmidli. Estimands and Their Role in Clinical Trials. *Statistics in Biopharmaceutical Research*, 9(3):268–271, jul 2017. ISSN 1946-6315. doi: 10.1080/19466315.2017.1302358.

[10] Craig Mallinckrodt, Geert Molenberghs, and Suchitrita Rathmann. Choosing estimands in clinical trials with missing data. *Pharmaceutical Statistics*, (August 2016), 2016. ISSN 15391612. doi: 10.1002/pst.1765.

[11] Siobhan Everson-Stewart and Scott S Emerson. Bio-creep in non-inferiority clinical trials. *Statistics in medicine*, 29(27):2769–2780, 2010.

[12] B. L Wiens and W. Zhao. The role of intention to treat in analysis of noninferiority studies. *Clinical Trials*, 4(3):286–291, jun 2007. doi: 10.1177/1740774507079443. URL http://dx.doi.org/10.1177/1740774507079443.

[13] Silvio Garattini et al. Non-inferiority trials are unethical because they disregard patients' interests. *The Lancet*, 370(9602):1875–1877, 2007.

[14] Brooke A. Rabe, Simon Day, Mallorie H. Fiero, and Melanie L. Bell. Missing data handling in non-inferiority and equivalence trials: A systematic review. *Pharmaceutical Statistics*, (April):1–12, 2018. ISSN 15391612. doi: 10.1002/pst.1867.

[15] National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials Panel on Handling Missing Data in Clinical Trials ; National Research*. 2010. ISBN 9780309158145.

[16] Angela M Wood, Ian R White, and Simon G Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4):368–376, jul 2004. doi: 10.1191/1740774504cn032oa. URL http://dx.doi.org/10.1191/1740774504cn032oa.

[17] Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in rcts; a review of the top medical journals. *BMC Med Res Methodol*, 14(1):118, 2014. doi: 10.1186/1471-2288-14-118. URL http://dx.doi.org/10.1186/1471-2288-14-118.

[18] Craig H. Mallinckrodt, W. Scott Clark, Raymond J. Carroll, and Geert Molenberghs. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics*, 13(2):179–190, 2003. ISSN 10543406. doi: 10.1081/BIP-120019265.

[19] Melanie L. Bell, Michael G. Kenward, Diane L. Fairclough, and Nicholas J. Horton. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ (Clinical research ed.)*, 346(jan21 1):e8668, 2013. ISSN 1756-1833. doi: 10.1136/bmj.e8668. URL http://www.bmj.com/cgi/doi/10.1136/bmj.e8668{\%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/23338004.

[20] JR Carpenter and MG Kenward. Missing data in randomised controlled trials - a practical guide. 2007. URL http://www.hta.nhs.uk/nihrmethodology/reports/1589.pdf.

[21] M. Matilde Sanchez and Xun Chen. Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Statistics in Medicine*, 25(7):1169–1181, 2006. doi: 10.1002/sim.2244. URL http://dx.doi.org/10.1002/sim.2244.

[22] Bongin Yoo. Impact of Missing Data on Type 1 Error Rates in Non-inferiority Trials. *Pharmaceutical Statistics*, 9(2):87–99, apr 2010. ISSN 15391604. doi: 10.1002/pst.378.

[23] Ilya Lipkovich and Brian L. Wiens. The role of multiple imputation in non-inferiority trials for binary outcomes. *Statistics in Biopharmaceutical Research*, 6315(November): 0–0, 2017. ISSN 1946-6315. doi: 10.1080/19466315.2017.1379433. URL https://www.tandfonline.com/doi/full/10.1080/19466315.2017.1379433.

[24] European Medicines Agency. Point to consider on switching between superiority and non-inferiority. 2000. URL http://www.eudra.org/emea.html.

[25] Petra Schiller, Nicole Burchardi, Michael Niestroj, and Meinhard Kieser. Quality of reporting of clinical non-inferiority and equivalence randomised trials - update and extension. *Trials*, 13(1), nov 2012. doi: 10.1186/1745-6215-13-214. URL http://dx.doi.org/10.1186/1745-6215-13-214.

[26] Erica Brittain and Daphne Lin. A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Statistics in Medicine*, 24(1):1–10, 2004. doi: 10.1002/sim.1934. URL https://doi.org/10.1002/sim.1934.

[27] Meg A Gamalo, Ram C Tiwari, and Lisa M Lavange. Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. 2013. doi: 10.1002/pst.1588.

[28] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G. Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychoudhury, and Laura Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014. ISSN 15391612. doi: 10.1002/pst.1589.

[29] Victoria Liublinska and Donald B. Rubin. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in Medicine*, 33(24):4170–4185, 2014. ISSN 10970258. doi: 10.1002/sim.6197.

[30] Mimi Kim, Cuiling Wang, and Xiaonan Xue. Assessing the influence of treatment nonadherence on noninferiority trials using the tipping point approach. *Statistics in Medicine*, (August 2018):650–659, 2018. ISSN 10970258. doi: 10.1002/sim.7999.

[31] Mallorie H. Fiero, Shuang Huang, Eyal Oren, and Melanie L. Bell. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*, 17(1), feb 2016. doi: 10.1186/s13063-016-1201-z. URL https://doi.org/10.1186/s13063-016-1201-z.

[32] Kenneth F Schulz, Douglas G Altman, and David Moher. Open Access CORRESPONDENCE BioMed Central CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. 8:18, 2010. URL http://www.biomedcentral.com/1741-7015/8/18.

[33] Ralph B. D'Agostino, Joseph M. Massaro, and Lisa M. Sullivan. Non-inferiority trials: Design concepts and issues - the encounters of academic consultants in statistics. *Statistics in Medicine*, 22(2):169–186, 2003. ISSN 02776715. doi: 10.1002/sim.1425.

[34] Steven A. Julious. Sample sizes for clinical trials with Normal data. *Statistics in Medicine*, 23(12):1921–1986, jun 2004. ISSN 0277-6715. doi: 10.1002/sim.1783. URL http://doi.wiley.com/10.1002/sim.1783.

[35] J L Schafer. Multiple imputation: a primer. *Stat Methods Med Res*, 8(1):3–15, 1999. ISSN 09622802. doi: 10.1191/096228099671525676.

[36] James R Carpenter, James H Roger, and Michael G Kenward. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(3):1352–71, 2013. ISSN 1520-5711. doi: 10.1080/10543406.2013.834911.

[37] Yang Yuan. Sensitivity Analysis in Multiple Imputation for Missing Data. *SAS Institute Inc*, pages 1–12, 2014.

[38] Ana Maria Lopez, Sandhya Pruthi, Judy C. Boughey, Marjorie Perloff, Chiu Hsieh Hsu, Julie E. Lang, Michele Ley, Denise Frank, Josephine A. Taverna, and H. H. Sherry Chow. Double-blind, randomized trial of alternative letrozole dosing regimens in postmenopausal women with increased breast cancer risk. *Cancer Prevention Research*, 9(2):142–148, 2016. ISSN 19406215. doi: 10.1158/1940-6207.CAPR-15-0322.

[39] Gilda Piaggio, Diana R. Elbourne, Stuart J. Pocock, Stephen J. W. Evans, Douglas G. Altman, and for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. *JAMA*, 308(24):2594, dec 2012. doi: 10.1001/jama.2012.87802. URL `http://dx.doi.org/10.1001/jama.2012.87802`.

[40] Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

[41] Roderick J.A. Little. Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250, 1982. ISSN 1537274X. doi: 10.1080/01621459.1982.10477792.

[42] G. Molenberghs, B. Michiels, M. G. Kenward, and P. J. Diggle. Monotone missing data and pattern-mixture models. *Stat. Neerl.*, 52(2):153–161, 1998. ISSN 0039-0402. doi: 10.1111/1467-9574.00075. URL `http://doi.wiley.com/10.1111/1467-9574.00075`.

[43] Craig K Enders. *Applied missing data analysis*. Guilford Press, 2010.

[44] Michael G. Kenward and James R. Carpenter. Multiple imputation. In *Longitudinal data analysis*, pages 477–499. Chapman and Hall/CRC, aug 2009. doi: 10.1201/9781420011579-33. URL `https://www.taylorfrancis.com/books/e/9781420011579/chapters/10.1201/9781420011579-33`.

[45] Ian R. White, James Carpenter, Stephen Evans, and Sara Schroter. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clinical Trials*, 2007. ISSN 17407745. doi: 10.1177/1740774507077849.

[46] Donald B Rubin. Multiple imputation for nonresponse in surveys (wiley series in probability and statistics). 1987.

[47] Y C Yuan. P267-25: Multiple imputation for missing data: concepts and new development. *SAS White Papers*, pages 1–13, 2000.

[48] Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of monte carlo methods*, volume 706. John Wiley & Sons, 2013.

[49] Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. (November 2017), 2018.

[50] Melanie L Bell, Lysbeth Floden, Brooke A Rabe, Stacie Hudgens, Haryana M Dhillon, Victoria J Bray, and Janette L Vardy. Analytical approaches and estimands to take account of missing patient-reported data in longitudinal studies. *Patient related outcome measures*, 10:129, 2019.

[51] Finbarr P Leacy, Sian Floyd, Tom A Yates, and Ian R White. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. 185(4):304–315, 2017. ISSN 14766256. doi: 10.1093/aje/kww107.

[52] Thomas Permutt. A taxonomy of estimands for regulatory clinical trials with discontinuations. *Statistics in Medicine*, 35(17):2865–2875, 2016. ISSN 10970258. doi: 10.1002/sim.6841.

Appendix A

Systematic Review of Non-Inferiority Trials

Received: 14 August 2017 | Revised: 4 April 2018 | Accepted: 10 April 2018

DOI: 10.1002/pst.1867

MAIN PAPER

WILEY

Missing data handling in non-inferiority and equivalence trials: A systematic review

Brooke A. Rabe[1] | Simon Day[2] | Mallorie H. Fiero[3] | Melanie L. Bell[4]

[1]Interdisciplinary Program in Statistics, The University of Arizona, Tucson, AZ, USA

[2]Clinical Trials Consulting & Training Limited, UK

[3]Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

[4]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, The University of Arizona, Tucson, AZ, USA

Correspondence
Brooke A. Rabe, The University of Arizona, Interdisciplinary Program in Statistics, 617 N. Santa Rita Ave, Tucson, AZ 85721, USA.
Email: brooker@email.arizona.edu

Summary

Background: Non-inferiority (NI) and equivalence clinical trials test whether a new treatment is therapeutically no worse than, or equivalent to, an existing standard of care. Missing data in clinical trials have been shown to reduce statistical power and potentially bias estimates of effect size; however, in NI and equivalence trials, they present additional issues. For instance, they may decrease sensitivity to differences between treatment groups and bias toward the alternative hypothesis of NI (or equivalence).

Aims: Our primary aim was to review the extent of and methods for handling missing data (model-based methods, single imputation, multiple imputation, complete case), the analysis sets used (Intention-To-Treat, Per-Protocol, or both), and whether sensitivity analyses were used to explore departures from assumptions about the missing data.

Methods: We conducted a systematic review of NI and equivalence trials published between May 2015 and April 2016 by searching the PubMed database. Articles were reviewed primarily by 2 reviewers, with 6 articles reviewed by both reviewers to establish consensus.

Results: Of 109 selected articles, 93% reported some missing data in the primary outcome. Among those, 50% reported complete case analysis, and 28% reported single imputation approaches for handling missing data. Only 32% reported conducting analyses of both intention-to-treat and per-protocol populations. Only 11% conducted any sensitivity analyses to test assumptions with respect to missing data.

Conclusion: Missing data are common in NI and equivalence trials, and they are often handled by methods which may bias estimates and lead to incorrect conclusions.

1 | INTRODUCTION

Non-inferiority (NI) clinical trials are designed to test whether a new treatment is not unacceptably worse than an existing treatment. In statistical terms, an NI trial tests the null hypothesis that a new treatment is less effective by a pre-defined NI margin, δ, than a treatment that has already been shown to be superior to placebo.[1-3] Equivalence trials, which are less common than NI trials, aim to show 2 treatments are therapeutically similar within some pre-defined equivalence margin.[4] These designs are commonly used to compare efficacy of a new treatment to an

existing standard of care when the new treatment is believed to have other advantages that may be related to cost, convenience of delivery, improved safety, fewer side effects, etc..[3,5] The design and analysis of NI and equivalence trials are generally more challenging than superiority studies: the choice of a margin, δ, and issues arising from absence of a placebo arm (which is not usually included) must be considered carefully to avoid invalid conclusions.[3,6] Furthermore, the interpretation of NI and equivalence trials is complicated by several factors, including that of missing data.[3,6,7]

In the missing data framework developed by Rubin, there are 3 types of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).[8] Data MCAR are a relatively rare and benign form: the missing data are not related to any other data, observed or unobserved. Data MAR may be related to observed data but, conditioned on these data, data MAR are then MCAR. Data MNAR are related to unobserved data—the missing value is related to the reason it is missing. It is worth noting that every approach to missing data handling makes assumptions about the underlying mechanism of missingness and about the full data model. However, methods for handling data MNAR (and MAR) must extrapolate from observed values, requiring additional assumptions which are untestable and unverifiable.[9]

Statistically based approaches to handling missing data, such as multiple imputation[10] and weighted generalized estimating equations (GEE),[11] try to make reasonable and justifiable assumptions about the underlying missingness mechanism, often assuming data are MAR.[12] Complete-case analysis, in which cases with missing values are excluded from analysis, and single imputation, where a single value such as the last observation or the worst-case outcome replaces the missing value, are simple to implement; however, these approaches often inadequately account for the pattern of missingness and fail to reflect the true uncertainty in reported estimates.[12] To assess the robustness of conclusions, sensitivity analyses that explore deviations from the primary missingness assumptions are recommended.[9]

Missing data have been widely investigated in the superiority trial context. They often lead to reduced power and biased estimates, particularly when the proportion of missing data is high. While missing data in NI and equivalence trials present many similar challenges, they demand additional considerations. For instance, in superiority studies, it is generally believed any bias arising from missing values is conservative in nature, favoring the null hypothesis of no difference by decreasing sensitivity to differences between treatment groups. In NI trials, however, missing data may bias results toward the alternative hypothesis of NI.[13]

Further complicating the interpretation of NI trials is the choice of an appropriate analysis set, a key element in defining the estimand, which, based on the trial objectives, describes what is being estimated and in which population.[14,15] In superiority studies, whenever participants drop out or fail to comply with treatment protocols, the principle of intention-to-treat (ITT) is usually applied, ensuring groups are analyzed as randomized regardless of adherence to treatment. Conclusions are deemed conservative because they often bias results toward the null hypothesis of no difference. In a NI setting, however, an ITT analysis which does not thoughtfully impute missing data may be anti-conservative.[3,7] On the other hand, per-protocol (PP) analyses, in which non-adherent individuals and those for whom data are missing for the primary estimand are not included in the analysis, are not necessarily preferable. Simulation studies show that the conservatism or anti-conservatism of the analysis depends on various factors, among them, the type of missingness and the methods for handling missing data in the ITT analysis set.[16,17] General recommendations for NI trials are to include both ITT and PP analyses.[3] Although the ITT and PP analyses are commonly believed to represent extremes, both may be biased in the same direction.[7] This highlights the need for sensitivity analyses which vary assumptions about missingness according to the chosen estimands and the analysis sets used.

In NI and equivalence trials, assumptions of assay sensitivity and constancy of effect are critical, yet both may be invalidated in the presence of missing data.[18] Assay sensitivity, which is a trial's ability to detect differences of a specified size between treatments,[3] is necessarily demonstrated in superiority studies if the investigational treatment is shown to be superior to the control (whether active treatment or placebo). However, showing a treatment is non-inferior does not demonstrate assay sensitivity. The constancy assumption helps to support assay sensitivity by holding that the size of the treatment effect of the active control (AC) compared with placebo in historical trials would be maintained in the NI trial. If missing data are handled differently than in the historical studies, the constancy assumption may be violated.[18] In theory, this means that if the NI trial had included a placebo arm, due to the different methods of missing data handling, the AC could appear to perform no better than placebo in the NI trial. Without direct comparison against a concurrent placebo arm, the analysis could result in falsely concluding a new treatment is effective (non-inferior or equivalent) when, in fact, the AC itself would appear no better than placebo.

Since 2007, the number of published NI trials has vastly increased.[5] Although the extension of the CONSORT 2010 Statement on reporting of NI and equivalence trials has likely improved the quality of reporting, this guidance does not address missing data.[4] Conversely, the National Research Council's 2010 report on missing data in clinical trials does not address NI trials.[9] Neither the ICH E9[1] nor the European Medicines Agency[19,20] provide guidance on missing data handling in the NI trial context. Several systematic reviews of NI trials have investigated methodological approaches and quality of reporting. These reviews show that even in high impact journals, there remain problems with reporting on the methodological approach, interpretation of the result of NI, and on the choice of the NI margin.[21-23] Meanwhile, reviews of missing data in randomized controlled trials (RCTs) report that missing data are common, and that simple methods for handling missing data, although lacking in justification, remain popular.[24,25] Furthermore, contrary to National Research Council recommendations, typically very few sensitivity analyses are carried out.[24-26] A recent review has addressed missing data in NI trials, but the extent of missing data found was not reported.[22]

The primary aim of this study was to review in NI (and equivalence) trials the extent, handling, and use of sensitivity analyses for missing data and the choice of population sets in the primary analysis: PP, ITT, or both. Our secondary aim was to describe general characteristics of the trials such as features of the design, and quality of reporting with respect to missing data, with specific aims focusing on reporting on the NI margin, sample size calculation, discussion of the constancy assumption and assay sensitivity, and comparison of reporting quality between journals with high and low impact factors.

2 | METHODS

We conducted this systematic review in accordance with PRISMA statement guidelines.[27] We searched the PubMed database for NI and equivalence trials published in English between May 1, 2015 and April 30, 2016. Within the title or abstract, we required "randomized (randomised)" and at least 1 of the following keywords: "non-inferiority (noninferiority)", "active control", "equivalence", "equivalency", or "statistical(ly) equivalent". The choice of keywords was based on search terms used in other systematic reviews of NI and equivalence trials.[21,28] We included RCTs with either an explicit statement concerning the NI margin or the intention to conduct an NI (or equivalence) trial. We excluded studies testing bioequivalence, protocols of trials, pilot studies, observational studies, secondary reports, and trials with survival outcomes. The statistical issues arising from missing data are different in trials with survival outcomes as patients who drop out are often considered censored in time-to-event analyses.[24] To compare our results with prior reviews that focused on the top medical journals, we oversampled from Annals of Internal Medicine, BMJ, JAMA, The Lancet, and The New England Journal of Medicine. To increase generalizability, we sampled from all other journals regardless of impact factor. After article selection, we broadened our definition of high impact to include all journals with impact factor greater than 5.

The search, screening, and most of the content assessment was performed by 1 reviewer (B.R.). Twenty-three articles were reviewed by a second reviewer (M.F.). To determine inter-reviewer agreement, both reviewers extracted data from 6 articles and values were compared. All discrepancies were minor and were discussed in consultation with a third party (M.B.). An additional 3 articles were reviewed by both (B.R., M.F.) for initial training.

2.1 | Data extraction

Data collected included the number and proportion of participants with missing primary outcome data. For primary outcomes measured repeatedly, we used the final follow-up timepoint to calculate the missing proportion, unless a different time point was specified for the primary analysis. Participants who dropped out of trials and who were then defined as non-responders were counted as missing data. The quantity of missing data was determined from the CONSORT flow diagram if available and from a thorough reading of the results. We extracted the methods for the primary analysis, as well as methods for missing data handling in the primary analysis: complete case, single imputation (including worst case, best case, mean imputation, regression imputation, last observation carried forward, or baseline observation carried forward), multiple imputation, mixed model (which uses repeated measures on the primary outcome variable), GEE, unclear, etc. We considered the missing data handling to be by a complete cases analysis if no imputation was described and if the authors did not use a model-based approach, such as a mixed model, in the primary analysis. We recorded whether a sensitivity analysis for missing data was done and details about the methods used and assumptions. If sensitivity analysis was done for both ITT and PP analysis sets, we recorded whether the analysis was

different between the 2 sets. We recorded the analysis sets used (ITT, modified ITT, PP, or both), using the authors' definitions of these populations. In general, outcome data from patients that were lost-to-follow-up, withdrawn from the trial, or excluded from the primary analysis for protocol violations were considered missing. We classified any analysis set labeled "as-treated" as PP.

We collected data on the type of trial (drug, device, other), type of outcome (binary, continuous, count), and sample size which we defined as the number of randomized participants. We reported whether the studies had multiple primary outcomes, and if so, whether one was for testing superiority and the other for NI. If a sample size calculation was presented, we recorded whether the sample size was increased to compensate for any potential missing outcome data and whether the NI (or equivalence) design was accounted for. We reported how often the primary outcome measures were collected (repeated or single) and how outcomes were treated in the primary analysis (repeated or single). This was to determine whether the analyses benefited from repeated collection of outcomes if these data were collected.[29] We also searched for any reference to the constancy assumption, assay sensitivity or reporting of a trial comparing the AC to placebo, and whether there was discussion of possible violations of the constancy assumption with reference to missing data. We reported whether the authors justified the NI or equivalence margin. We used lenient criteria to determine whether a margin was justified: at minimum, we looked for a reference to regulatory guidelines, or to prior, similar studies which used the same margin. Finally, we collected data on study conclusions, the journal, its impact factor, and the study sponsor or source of funding.

3 | RESULTS

On June 27, 2016, the PubMed database search returned 1813 articles, 47 of which were from the journals Annals of Internal Medicine, BMJ, JAMA, The Lancet, and The New England Journal of Medicine. All these articles were screened and ultimately 26 were reviewed. From the other journals, 93 were selected randomly for screening, and 10 were excluded, leaving 83 to be reviewed. In total, 109 articles were reviewed. Figure 1 shows details of the study selection.

General characteristics of the articles reviewed are shown in Table 1. Articles from journals identified as high impact (impact factor score greater than 5) comprised 47% of the sample. Most studies were funded by industry (48%) or government (28%). Sixty-three percent were drug trials. Medical specialty fields were diverse and included oncology, cardiology, infectious disease and vaccines, diabetes, and other internal medicine. Eighty-nine percent were NI trials; 11% were equivalence trials.

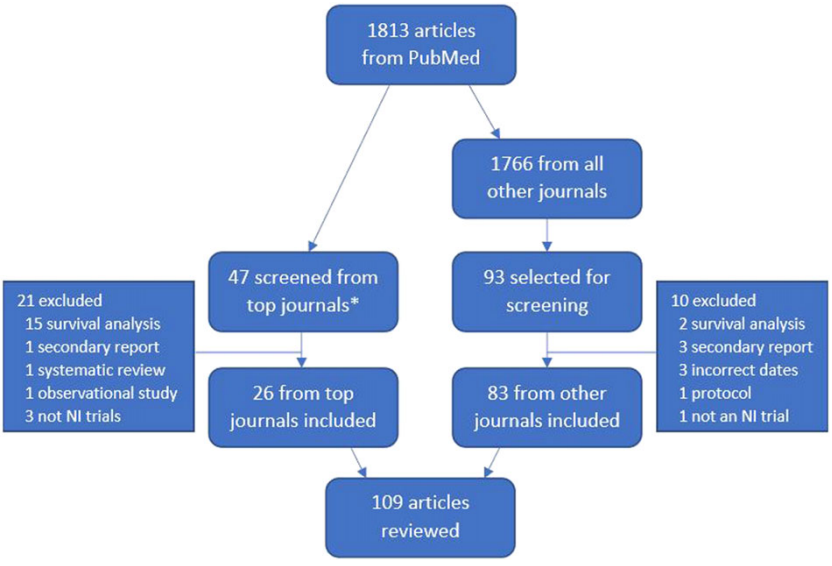

FIGURE 1 Flow diagram depicting article selection. *top medical journals initially identified as Annals of Internal Medicine, BMJ, JAMA, The Lancet, New England Journal of Medicine

TABLE 1 Trial characteristics conditioned on proportion of missing data in primary outcome

	<10% Missing (*N* = 60) *N* (%)	≥10% Missing (*N* = 49) *N* (%)	Total (*N* = 109) *N* (%)
Journal impact factor			
Impact factor > 10	20 (33)	14 (29)	34 (31)
5 < impact factor < 10	9 (15)	8 (16)	17 (16)
Impact factor < 5	31(52)	27 (55)	58 (53)
Funding			
Pharma, device, or other industry	33 (55)	19 (39)	52 (48)
Government	12 (20)	19 (39)	31 (28)
Non-profit	4 (7)	4 (8)	8 (7)
Combination	4 (7)	4 (8)	8 (7)
None	7 (12)	3 (6)	10 (9)
Trial type			
Drug	39 (65)	30 (61)	69 (63)
Device	8 (13)	6 (12)	14 (13)
Other	13 (22)	13 (27)	26 (24)
Field			
Cancer	8 (13)	4 (8)	12 (11)
Cardiology	9 (15)	4 (8)	13 (12)
Infectious disease and vaccines	11 (18)	14 (29)	25 (23)
Diabetes	5 (8)	3 (7)	8 (7)
Surgery and anesthesia	6 (10)	0 (0)	6 (6)
Dermatology	3 (5)	1 (2)	4 (4)
Mental health	0 (0)	2 (4)	2 (2)
Pediatrics	0 (0)	4 (8)	4 (4)
Obstetrics, gynecology, and urology	6 (10)	1 (2)	7 (6)
Other internal medicine	10 (17)	9 (18)	19 (17)
Other	2 (3)	7 (14)	9 (8)
Equivalence trials	6 (10)	6 (12)	12 (11)
Non-inferiority trials	54 (90)	43 (88)	97 (89)
Type of outcome			
Binary	43 (72)	27 (55)	70 (64)
Continuous	16 (27)	22 (45)	38 (35)
Count	1 (2)	0 (0)	1 (1)
How often outcome was collected			
Single	29 (48)	18 (37)	47 (43)
Repeated	31 (52)	31 (63)	62 (57)
How outcome was treated in the primary analysis			
Single	56 (93)	42 (86)	98 (90)
Repeated	4 (7)	7 (14)	11 (10)
Switched from superiority design	0 (0)	1 (2)	1 (1)
Multiple co-primary outcomes	10 (17)	6 (12)	16 (15)

(Continues)

TABLE 1 (Continued)

	<10% Missing (*N* = 60) *N* (%)	≥10% Missing (*N* = 49) *N* (%)	Total (*N* = 109) *N* (%)
One outcome superiority and another NI	2 (3)	2 (4)	4 (4)
NI margin justified	22 (37)	19 (39)	41 (38)
Presented sample size calculation	52 (87)	41 (84)	93 (85)
Considered missing data for sample size	34 (57)	25 (51)	59 (54)
Accounted for NI design in sample size	52 (87)	40 (82)	92 (99)
Reported reason why missing[a]	49 (92)	42 (88)	91 (90)

[a]Percentage calculated of 101 articles with missing primary outcomes.

3.1 | Design and quality of reporting

Table 1 also summarizes the general characteristics of trial design and quality of reporting. Binary primary outcomes were nearly twice as common as continuous outcomes. Sixty-two (62) percent did not justify the choice of NI or equivalence margin. Overall, 36 articles used a margin of either 10 or 15 percentage points, and only 8 justified this choice. While 57% of trials reported repeated collection of outcomes, among those only 18% treated the outcome as repeated in the primary analysis.

Sample size ranged from 26 to 2008 with a median size of 298 participants. Most articles showed a sample size calculation (85%); however, among those that did, 37% did not account for missing data. Nearly all accounted for the NI or equivalence design in the sample size calculation. One paper reported switching from superiority to a NI design as a better framework for addressing scientific aims. It was reported that the change was made before viewing any data or conducting analyses; the sample size calculation accounted for the NI design.[30]

3.2 | Extent and handling of missing data

The extent and handling of missing data are summarized in Table 2. Of the papers reviewed, 93% (101/109) were determined to have some missing primary outcome data. The mean proportion of missing outcome data was 10.6% [95% CI (8.9%,12.3%)], the median was 9.6%, and the maximum was 45.3%.

Among articles with missing data, for the primary analysis, 50% used a complete case analysis, and 28% used a single imputation approach such as last observation carried forward, worst-case, or best-case. Only 12% used multiple imputation or a model-based approach. The data were analyzed by the ITT principle in 71% of articles, and 61% reported conducting a PP analysis; 32% of articles reported that they conducted analyses using both ITT and PP analysis sets.

Sensitivity analyses with respect to missing data assumptions were reported in 11 articles (11% of those with any missing data). One trial explicitly referenced a MNAR model, but the details were only specified in supplemental materials.[31] One trial used a tipping-point analysis assuming missing data varied from worst to best case, but did not implement this in a multiple imputation framework, as advised by Yuan.[32] Seven of the sensitivity analyses used worst case imputation, and 2 used multiple imputation. Two of the sensitivity analyses made weaker assumptions about the missing mechanism than the primary analyses: 1 study used baseline-observation-carried-forward in the primary and multiple imputation in the sensitivity analysis (MCAR → MAR); the other study used multiple imputation in the primary and an unspecified MNAR sensitivity analysis (MAR → MNAR). Among the studies that analyzed both the ITT and PP populations in the primary analysis, none reported performing different sensitivity analyses for the different analysis sets. Of 6 trials with co-primary endpoints where one was superiority and the other was NI, only 1 paper used different sensitivity analyses to test assumptions about missing data for each primary endpoint. Some authors reported secondary analyses by redefining the analysis sets used (ITT or PP); these were not counted as sensitivity analysis in our review.

Among the 49 articles that reported more than 10% missing data, complete case analysis was used in 19 (39%), and only one of these did a sensitivity analysis. Five of the 51 studies that performed a complete case analysis also performed sensitivity analyses.

Two papers discussed the impact of missing data on the constancy assumption, and 2 others discussed how missing data may compromise assay sensitivity. Four of the 11 sensitivity analyses were reported in just 1 sentence, or not at all. The remaining 7 were reported in a paragraph, table, or in a detailed appendix.

TABLE 2 Missing data handling in the primary analysis conditioned on proportion of missing data

	<10% missing (*N* = 60) *N* (%)	≥10% missing (*N* = 49) *N* (%)	Total (*N* = 109) *N* (%)
Proportion of missing data in primary outcome			
0%-10%	60 (100)	--	60 (55)
10%-20%	--	34 (69)	34 (31)
20%-30%	--	9 (18)	9 (8)
Over 30%	--	6 (12)	6 (6)
Analysis sets			
ITT or modified ITT only	24 (40)	18 (38)	42 (39)
PP only	19 (32)	13 (27)	32 (29)
Both	17 (28)	18 (38)	35 (32)
Method[a]			
Complete case	32 (62)	19 (39)	51 (50)
Single imputation (LOCF, WC, BC, Reg., BOCF[b])	14 (27)	14 (29)	28 (28)
Unweighted GEE	1 (2)	0 (0)	1 (1)
Multiple imputation	0 (0)	6 (12)	6 (6)
Mixed model	1 (2)	5 (10)	6 (6)
Unclear	4 (8)	5 (10)	9 (9)
Sensitivity analysis reported[a]	5 (10)	6 (12)	11 (11)

[a]Percentage calculated of 101 articles with missing primary outcomes.

[b]Last observation carried forward, worst case, best case, regression, baseline observation carried forward.

We found that sensitivity analyses for missing data assumptions were significantly more likely to be reported in high impact journals (19% high impact versus 4% low impact, χ^2 *P*-value = 0.036). The proportion of articles reporting both ITT and PP analyses was greater in the high impact journals than in the low impact journals, but the difference was not statistically significant (37% high impact versus 28% low impact, χ^2 *P*-value = 0.38). To explore the sensitivity of these results to the choice of impact factor cutoff, we repeated these tests by redefining high impact journals as those with impact factor greater than 10. The results were consistent with those of the planned analyses. Overall, 90 of 109 (83%) trials concluded NI (or equivalence) of the experimental treatment to the control. Table 3 shows sensitivity analyses, analysis sets, and amount of missing data stratified by conclusion. No significant differences were found between articles that did and did not conclude NI.

Comparison of our primary results to other reviews is shown in Table 4. The proportion of trials with missing data and the approaches used were consistent with findings of other systematic reviews of missing data in RCTs; however, the frequency of sensitivity analyses (11%) for testing missing data assumptions was lower than reported in other reviews (16%-37%). Compared with Bell et al, who reviewed trials in leading medical journals, the percentage of papers reporting sensitivity analyses was significantly lower ($P < 0.001$) in a test for difference in proportions with continuity correction; however, compared with Fiero et al, no significant difference was found ($P = 0.288$). Rehal et al only very broadly defined sensitivity analyses (not specific to missing data assumptions); therefore, no formal test comparing these results was performed.

4 | DISCUSSION

4.1 | Summary

Our systematic review of 109 NI and equivalence trials sampled from both high and low impact factor journals found missing data were reported in 93% of trials, a proportion which is consistent with previous reviews of missing data in other types of RCTs.[24-26] Among all trials, the mean proportion of missing primary outcome data was 10.6%, and 45%

TABLE 3 Selected results stratified by conclusion

	Conclude Noninferiority (or Equivalence)	
	Yes ($N = 90$) N (%)	No[a] ($N = 19$) N (%)
Sensitivity analysis reported	8/83[b] (10)	3/18[b] (17)
Analysis sets		
ITT or modified ITT only	35 (39)	7 (37)
PP only	28 (31)	4 (21)
Both	27 (30)	8 (42)
Percent missing data		
0%-10%	51 (57)	9 (47)
10%-20%	29 (32)	5 (26)
20%-30%	5 (6)	4 (21)
Over 30%	5 (6)	1 (5)

[a]Includes 1 article reporting inconclusive results.

[b]Articles with missing data.

TABLE 4 Comparison of systematic reviews. With the exception of Rehal et al, 2016, percentages calculated for missing data approaches and sensitivity analyses are based on the number of papers with missing data. All percentages in Rehal et al use total number of papers reviewed in the denominator because the number of papers with missing data was not reported

Study	Type	Number of Trials	Number (%) Papers with Missing Data	Number (%) Papers with > 10% Missing	Missing Data Approaches	Number (%)	Number (%) Reporting Sensitivity Analyses
Bell et al, 2014	Randomized controlled trials (top medical journals)	77	73 (95)	36 (47)	Complete case Single imputation Multiple imputation Model based	33 (45) 20 (27) 6 (8) 14 (19)	27 (37)
Fiero et al, 2016	Cluster randomized trials (all journals)	86	80 (93)	58 (67)	Complete case Single imputation Multiple imputation Model based	44 (55) 6 (8) 2 (3) 18 (23)	14 (18)
Rehal et al, 2016	Non-inferiority trials (top medical journals)	168	Unreported	Unreported	Complete case/no imputation/not reported Single imputation Multiple imputation Model based	117 (70) 35 (21) 11 (7) Unclear	27 (16)
Current study	Non-inferiority trials (all journals)	109	101 (93)	49 (45)	Complete case Single imputation Multiple imputation Model based Unclear	51 (50) 28 (28) 6 (6) 6 (6) 9 (9)	11 (11)

reported more than 10% missing. The most common approaches to handling missing data were complete case analysis (50%) and single imputation (28%); however, articles reporting greater than 10% missing data were less likely to use complete case analysis ($P = 0.037$). Multiple imputation and model-based approaches were rare even when the proportion of missing data was high. Sensitivity analyses examining the impact of assumptions around the missingness mechanism were seen in only 11 papers, less often than expected based on past reviews.[22,24,25] Only 32% of articles reported performing primary analyses on both the ITT and PP population sets in accordance with most guidelines on NI trials.[3]

Researchers did not report making different assumptions about missingness mechanisms for the ITT and PP populations if both were analyzed. A majority (57%) of trials reported that the outcome measure (on which the primary outcome is based) was collected repeatedly; however, only 10% used a model that could account for repeated observations on participants. Sample size calculations often failed to account for missing data, and NI margins were usually presented without justification. Furthermore, very few authors discussed potential violations of the constancy assumption in NI trials. Articles published in high impact journals may be associated with higher quality reporting as is suggested by a greater likelihood of conducting sensitivity analyses with respect to missing data assumptions.

4.2 | Relation to other literature

In 2010, the National Research Council published guidance on reducing the frequency of missing data in clinical trials.[9] In comparing our review to other systematic reviews of RCTs, we do not find evidence of any improvements to data collection and trial design with the effect of reducing missing data. Previous systematic reviews investigating the extent of missing data in RCTs looked at the percent of trials with substantial missing data (>10%) and found this to be 27% (all RCTs in 1997),[33] 51% (all RCTs excluding survival outcomes in 2001),[26] and 47% (all RCTs excluding survival outcomes in 2013).[24] These reviews and others are consistent with our finding that most researchers do not deal with missing data adequately and fail to examine assumptions related to missingness through sensitivity analyses.[22,24,25]

Rehal et al examined quality of reporting in NI trials with an interest in how missing data were handled but did not report the proportion of trials with missing primary outcome data. They reported ITT analyses in 77% and PP analyses in 54% of papers reviewed.[22] These values are close to our own findings on analysis sets (71% for ITT/modified ITT, 61% for PP). In addition, Rehal et al reported a weak association between concluding NI and quality of reporting. Our review did not find that concluding NI was associated with sensitivity analyses or reporting analyses of both ITT and PP population sets. We obtained a slightly lower than expected frequency of sensitivity analyses than reported in reviews of other designs.[22,24-26] This could be due to our strict adherence to recording only those sensitivity analyses specifically related to missing data assumptions. Some authors used the term sensitivity analysis to refer to making changes to the analysis sets, for example, by changing the definitions of the ITT and PP definitions; however, these were not considered as these are secondary analyses and do not target questions about the validity of the missing data mechanism.

4.3 | Strengths and limitations

This review is the first to look at the extent of missing data in NI trials. We used a large sample of recent publications both from high and lower impact journals which give not only a clear picture of current practices among top researchers but also among typical researchers. In addition, we specifically looked at how assumptions about missing data were made vis-à-vis the analysis sets in sensitivity analyses if both ITT and PP sets were analyzed.

Our study was limited by the lack of a universally agreed definition of the ITT analysis set.[33,34] Our review considered ITT or modified ITT as defined by the authors, and we did not distinguish between those various definitions: some analysis sets were determined at randomization, some after initial follow-up, and some defined by first treatment. Many authors excluded data for various reasons (sometimes for not completing treatment, or for protocol violations) and called this the ITT analysis set. In our review, such a modified ITT analysis with no imputation was considered a complete-case analysis unless a statistically based approach, such as a mixed model, was used to handle missing data in the primary analysis. Hence, this ambiguity also affected how the missing-data handling method was defined.

We found very few references to problems arising from the constancy assumption or assay sensitivity with respect to missing data in NI trials. Some authors might have discussed these assumptions but not referenced them by name; therefore, it is possible that our review underestimated the number of studies discussing potential violations.

4.4 | Recommendations

In superiority trials, complete case analysis, last observation carried forward, and other single imputation methods for handling missing data often reduce statistical power to conclude superiority. However, in NI trials, approaches that assume data are MCAR have been shown to inflate type I error.[18] Researchers should improve control of type I error and mitigate bias by imputing under the assumption of the null hypothesis of inferiority as suggested by FDA guidelines.[3,7] Furthermore, it is difficult to know how missing data may critically affect NI trial conclusions without some sensitivity analyses that investigate alternative missingness assumptions. Tipping point analyses, which impute

values over a range of plausible scenarios, have also been proposed for NI trials, and we support this approach.[35] Prespecification of analysis plans, sensitivity analyses and the analysis sets, in addition to reporting that they were prespecified, would be useful. Our review found that most researchers did not distinguish between treatment discontinuation and dropout. As recommended by the NRC panel on missing data, trialists should make this distinction and report both treatment discontinuation and dropout, and they should continue to assess patients even after discontinuation.[9]

Data that have been collected repeatedly might be used more effectively in the primary analysis especially with missing data present, even while ensuring that the correct estimand is analyzed.[29] In our review, many trials reporting repeated collection of the primary outcome variable did not use these data to their full extent. This is a lost opportunity to possibly mitigate some bias introduced by missing data.

In NI trials, as in all clinical trials, analyses performed on the ITT and PP populations attempt to answer different questions about the study treatment: an ITT analysis may target the question of how the treatment fares under typical use or as a treatment *policy*, while a PP analysis aims to measure efficacy under full compliance. Many researchers are increasingly aware of the importance of well-defined estimands and how the definitions of the ITT and PP analysis sets are key components of these estimands. Furthermore, it should be noted performing and reporting both an ITT and a PP analysis may not be adequate for determining efficacy in the presence of missing data.[36] Robust sensitivity analyses are necessary, and the assumptions about missing data imputation should be clearly stated and be consistent with the primary estimands.[15] Carpenter and Kenward[12] write that if data are missing after deviation from protocol, then missing data handling methods must be different for ITT and PP assumptions. In the PP case, data are imputed under the assumptions based on continuation of protocol, while in the ITT case we might reasonably assume that data are observed if and only if participants comply, meaning they are implicitly MNAR.[12]

5 | CONCLUSION

Missing data remain a problem in NI and equivalence trials and methods for handling them are often not statistically based. Researchers should pay closer attention to the assumptions implicit in their methods for handling missing data and how these might support or contradict the estimands. More researchers should conduct analyses and report estimates for both ITT and PP populations. Sensitivity analyses should be performed and reported to add robustness to trial conclusions, especially when the proportion of missing data is high. Many researchers miss the opportunity to reach more robust conclusions with model-based approaches when data are collected repeatedly but only analyzed singly.

Current recommendations by academic panels and regulatory agencies are not being followed consistently. Researchers should prioritize collection of complete data, even after discontinuation of treatment or deviation from protocol, to improve study quality and minimize the chance of falsely concluding NI. The consequences of mishandling missing data in NI trials are serious: if a new treatment is incorrectly deemed non-inferior, patients may be put at risk.

DISCLAIMER

This article reflects the views of the authors and should not be construed to represent FDA's views or policies. This work was completed at the University of Arizona.

ORCID


Brooke A. Rabe http://orcid.org/0000-0003-3949-8494

REFERENCES

1. International Conference on Harmonisation. Statistical Principles for Clinical Trials E9. 1998. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf. Accessed July 18, 2017.

2. International Conference on Harmonisation. Choice of Control Group and Related Issues in Clinical Trials E10. 2000. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf. Accessed July 18, 2017.

3. US Food and Drug Administration. Non-inferiority clinical trials to establish effectiveness. Guidance for Industry. 2016. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. Accessed July 18, 2017.

4. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG. Reporting of noninferiority and equivalence randomized trials extension of the CONSORT 2010 statement reporting noninferiority and equivalence. *JAMA*. 2012;308(24):2594-2604. https://doi.org/10.1001/jama.2012.87802
5. Wangge G, Klungel OH, Roes KCB, De Boer A, Hoes AW, Knol MJ. Should non-inferiority drug trials be banned altogether? *Drug Discov Today*. 2013;18(11–12):601-604. https://doi.org/10.1016/j.drudis.2013.01.003
6. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Stat Med*. 2003;22(2):169-186. https://doi.org/10.1002/sim.1425
7. Wiens BL, Zhao W. The role of intention to treat in analysis of noninferiority studies. *Clin Trials*. 2007;4(3):286-291. https://doi.org/10.1177/1740774507079443
8. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. https://doi.org/10.1093/biomet/63.3.581
9. National Research Council Panel on Handling Missing Data in Clinical Trials. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington (DC): National Academies Press; 2010 https://doi.org/10.17226/12955.
10. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8(1):3-15. https://doi.org/10.1191/096228099671525676
11. Preisser JS, Lohman KK, Rathouz PJ. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Stat Med*. 2002;21(20):3035-3054. https://doi.org/10.1002/sim.1241
12. Carpenter J, Kenward M. Missing data in randomised controlled trials—a practical guide. 2007. Health Technology Assessment Methodology Programme, Birmingham, p. 199.
13. Fleming TR, Odem-Davis K, Rothmann MD, Li Shen Y. Some essential considerations in the design and conduct of non-inferiority trials. *Clin Trials*. 2011;8(4):432-439. https://doi.org/10.1177/1740774511410994
14. Akacha M, Bretz F, Ohlssen D, Rosenkranz G, Schmidli H. Estimands and their role in clinical trials. *Stat Biopharm Res*. 2017;9(3):268-271. https://doi.org/10.1080/19466315.2017.1302358
15. International Conference on Harmonization. E9 (R1) Estimands and sensitivity analysis in clinical trials. 2017. http://academy.gmp-compliance.org/guidemgr/files/E9-R1EWG_STEP2_GUIDELINE_2017_0616.pdf. Accessed January 9, 2018.
16. Sanchez M, Chen X. Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Stat Med*. 2006;25(7):1169-1181. https://doi.org/10.1002/sim.2244
17. Brittain E, Lin D. A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Stat Med*. 2005;24(1):1-10. https://doi.org/10.1002/sim.1934
18. Wiens BL, Rosenkranz GK. Missing data in non-inferiority trials. *Stat Biopharm Res*. 2013;6315(April 2015):37-41. https://doi.org/10.1080/19466315.2013.847383
19. Committee for Medicinal Products for Human Use (CHMP). Guideline on the choice of the non-inferiority margin. EMEA. 2005. http://www.emea.eu.int. Accessed August 10, 2017.
20. Committee for Proprietary Medicinal Project (CPMP). Points to consider on switching between superiority and non-inferiority. EMEA. 2000. http://www.eudra.org/emea.html. Accessed August 10, 2017.
21. Schiller P, Burchardi N, Niestroj M, Kieser M. Quality of reporting of clinical non-inferiority and equivalence randomised trials—update and extension. *Trials*. 2012;13(1):214. https://doi.org/10.1186/1745-6215-13-214
22. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PPJ. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open*. 2016;6(10):e012594. https://doi.org/10.1136/bmjopen-2016-012594
23. Wangge G, Klungel OH, Roes KCB, de Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. *PLoS One*. 2010;5(10):e13550. https://doi.org/10.1371/journal.pone.0013550
24. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):118. https://doi.org/10.1186/1471-2288-14-118
25. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*. 2016;17(1):72. https://doi.org/10.1186/s13063-016-1201-z
26. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1(4):368-376. https://doi.org/10.1191/1740774504cn032oa
27. Moher D, Liberati A, Tetzlaff J, Altman DG, Altman D. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097
28. Lange S, Freitag G. Special invited papers section: therapeutic equivalence—clinical issues and statistical methodology in noninferiority trials. *Biom J*. 2005;47(1):12-27. https://doi.org/10.1002/bimj.200410085
29. Ashbeck EL, Bell ML. Single time point comparisons in longitudinal randomized controlled trials: power and bias in the presence of missing data. *BMC Med Res Methodol*. 2016;16. https://doi.org/10.1186/s12874-016-0144-0.
30. Christensen KD, Roberts JS, Whitehouse PJ, et al. Disclosing pleiotropic effects during genetic risk assessment for alzheimer disease: a randomized trial. *Ann Intern Med*. 2016;164(3):155-163. https://doi.org/10.7326/m15-0187

31. Rahman NM, Pepperell J, Rehal S, et al. Effect of opioids vs NSAIDs and larger vs smaller chest tube size on pain control and pleurodesis efficacy among patients with malignant pleural effusion: the TIME1 randomized clinical trial. *JAMA*. 2015;314(24):2641-2653. https://doi.org/10.1001/jama.2015.16840

32. Yuan Y. Sensitivity analysis in multiple imputation for missing data. *SAS Inst Inc*. 2014;1-12.

33. Hollis S, Campbell F. What is meant by intention to treat analysis ? Survey of published randomised controlled trials. *BMJ*. 1999;319(7211):670-674. https://doi.org/10.1136/bmj.319.7211.670

34. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. *BMJ*. 2010;340(jun14 1):c2697. https://doi.org/10.1136/bmj.c2697

35. Yan X, Lee S, Li N. Missing data handling methods in medical device clinical trials. *J Biopharm Stat*. 2009;19(6):1085-1098. https://doi.org/10.1080/10543400903243009

36. Rosenkranz GK. Analysis sets and inference in clinical trials. *Ther Innov Regul Sci*. 2013;47(4):455-459. https://doi.org/10.1177/2168479013486270

How to cite this article: Rabe BA, Day S, Fiero MH, Bell ML. Missing data handling in non-inferiority and equivalence trials: A systematic review. *Pharmaceutical Statistics*. 2018;17:477–488. https://doi.org/10.1002/pst.1867