# New Methods and Software to Determine the Impact of Missing Data in Clinical Trials

Daniel Scharfstein, ScD; Aidan McDermott, PhD; Elizabeth Stuart, PhD;Tianjing Li, PhD; Chenguang Wang, PhD

**Awardee Institution:** John Hopkins University

**Original Project Title:** Sensitivity Analysis Tools for Clinical Trials with Missing Data

**HSRProj ID:** 20143591

**PCORI ID:** ME-1303-6016

To cite this document, please use: Scharfstein D, McDermott A, Stuart E, Li T, Wang C. (2019). *New Methods and Software to Determine the Impact of Missing Data in Clinical Trials.* Washington, DC: Patient-Centered Outcomes Research Institute (PCORI).
https://doi.org/10.25302/11.2019.ME.13036016

# Table of Contents



[1] Some material in this section as been reproduced, with permission, from [43]

[2] Some material in this section as been reproduced, with permission, from [43]

# B Abstract

**Background:** Missing outcome data are a widespread problem in randomized trials with repeated assessments, including those with patient-centered outcomes. The essential problem is that inference about treatment effects relies on unverifiable assumptions about the nature of the mechanism that generates the missing data, leading to concerns about the validity and robustness of trial results. To address this problem, the 2010 National Research Coun- cil (NRC) Report recommended that "examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting." PCORI Methodology Standard MD-5 echoes this recommendation. While Chapter 5 of the NRC Report
outlines a general framework for conducting sensitivity analysis (like "stress testing"), there are two major problems with existing methods that have limited their usefulness: (1) they have not been implemented in software packages and (2) they do not adequately address non-monotone (i.e., intermittent) missing data patterns. The NRC Report recognized "the development of software that supports coherent missing data analyses" as a "high priority" and highlighted non-monotone missing data as one of the "important areas in which progress is particularly needed." The PCORI funding announcement for this project specifically called for the "development of software to reduce barriers that inhibit the use of rigorous methods for handling missing data."
**Objectives:** The objectives of this project were to (a) create unified and coherent methods for global sensitivity analysis of clinical trials with monotone and non-monotone missing data, (b) develop free, open source and reproducible software in SAS and R to implement the methods, (c) demonstrate the methods and software using clinical trial data with patient- centered outcomes and (d) disseminate the methods and software.
**Results:** Novel statistical methods and software were developed. The methodological approach for handling intermittent missing data involves multiply imputing, using a novel assumption and associated matching technique, the minimal number of observations to create datasets with monotone missing data structures. The sensitivity analysis approach, applied to each of the monotone datasets, involves conducting analyses under a novel class of missing data assumptions, indexed by sensitivity analysis parameters that quantify smooth depar-

tures from a reasonable benchmark assumption. For each assumption in the class, estimates of treatment effects are averaged over the monotone datasets and uncertainty is quantified using bootstrapping techniques. The sensitivity of the benchmark inference about treatment effects is then evaluated by comparing it with the inferences under alternative assumptions within the class. Methodology for conducting sensitivity analysis to the assumption used to create monotone datasets was not developed. While not initially part of the original scope of work, sensitivity analysis methods for analyzing randomized trials with death and missing data and for analyzing tuberculosis trials with intermittent missing data were developed.

A project website (`www.missingdatamatters.org`) was created to facilitate distribution of software, documentation, datasets, case studies, information about short courses and webinars, presentation materials and journal articles. More than 325 individuals registered to download software from the website. Over the course of the project, seven short-courses, 10 oral presentations, three webinars, one video-taped lecture and one poster were presented. Three case studies were developed. The main project manuscript that describes sensitivity analysis for trials with monotone missing data has been accepted for publication in *Biometrics*. A translational version of this paper has been accepted for publication in *Statistical Methods in Medical Research*. Two ancillary papers have been accepted for publication. A book describing the entirety of the methods and software is in preparation.

**Conclusions:** Despite wide dissemination efforts and deep engagement with the FDA, up- take of the methods and software has been slower than anticipated. Until investigators are required by FDA, PCORI, NIH and journals to rigorously evaluate the robustness of trial results to missing data assumptions, adoption of our technology is likely to be slow. Once the incentives are in place, our tools will be ready for use.

# C Background [3]

Missing outcome data are a widespread problem in clinical trials, including those with patient-reported outcomes. Such outcomes require active engagement of patients. While encouraged, patients are not required to remain or provide data while on-study. As a result, missing data can be expected [37].

To understand the magnitude of this issue, we reviewed randomized trials reporting five major patient-reported outcomes (SF-36, SF-12, Patient Health Questionnaire-9, Kansas City Cardiomyopathy Questionnaire, Minnesota Living with Heart Failure Questionnaire) published in five leading general medical journals (*New England Journal of Medicine*, *Journal of the American Medical Association*, *Lancet*, *British Medical Journal*, *PLoS One*) between January 1, 2008, and January 31, 2017. We identified 145 studies, which are summarized in Table 1 of Scharfstein and McDemott (2017) [44]. There is large variation in the percentages of missing data, with 78.6% of studies reporting percentages greater than 10%, 43.4% greater than 20% and 24.8% greater than 30%. Fielding *et al.* (2008) [16] conducted a similar review of clinical trials reporting quality of life outcomes in four of these journals during 2005/6 and found a comparable distribution of missing data percentages. Given the quality of these journals, it is likely that the percentages reported in [44, 16] are an optimistic representation of percentages of missing data across the universe of clinical trials with patient-reported outcomes published in the medical literature.

Missing outcome data complicates the inferences that can be drawn about treatment effects. While unbiased estimates of treatment effects can be obtained from trials with no missing data, this is no longer true when data are missing on some patients. The essential problem is that inference about treatment effects relies on *unverifiable* assumptions about the nature of the mechanism that generates the missing data. While we may know the reasons for missing data, we do not know the distribution of outcomes for patients with missing data, how it compares with that of patients with observed data and whether differences in these distributions can be explained by the observed data.

It is widely recognized that the way to address the problem caused by missing outcome

[3] Some material in this section as been reproduced, with permission, from [44]

data is to posit varying assumptions about the missing data mechanism and evaluate how inference about treatment effects is affected by these assumptions. Such an approach is called "sensitivity analysis." A 2010 National Research Council (NRC) report entitled "The Prevention and Treatment of Missing Data in Clinical Trials" [25] and a follow-up manuscript published in the *New England Journal of Medicine* [27] recommends:

> Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.

Li *et al.* (2012) [23] echoed this recommendation (see Standard 8) in their PCORI sponsored report entitled "Minimal Standards in the Prevention and Handling of Missing Data in Observational and Experimental Patient Centered Outcomes Research."

Chapter 5 of the NRC report lays out a general framework for conducting "global" sensitivity analysis [38, 43, 35, 39, 11, 41]. In this framework, the robustness of study results is evaluated across a broad range of assumptions that include a reasonable benchmark assumption and a collection of additional assumptions that trend toward best and worst case assumptions. From such an analysis, it can be determined how much deviation from the benchmark assumption is required in order for the inferences to change. If the deviation is judged to be sufficiently far from the benchmark assumption, then greater credibility is lent to the benchmark analysis; if not, the benchmark analysis can be considered to be fragile. Some researchers have dubbed this approach "tipping point analysis" [49, 4]. This "global" approach is contrasted with "local" sensitivity analysis [47, 9, 45, 29], which only evaluates robustness in a small neighborhood around a benchmark assumption and "ad-hoc" sensitivity analysis, which just evaluates robustness under a few different assumptions. The advantage of the "global" approach is that it is much more comprehensive.

The vast majority of global sensitivity analysis methods have focused on studies with monotone missing data (i.e., patients provide no data after first missed visit). Positing reasonable benchmark assumptions and an associated global sensitivity analysis methodology for studies with non-monotone missing data (i.e., patients provide data irregularly) is much more challenging. The NRC report highlighted non-monotone missing data as one of the

"important areas in which progress is particularly needed."

Global sensitivity analyses are rarely reported. There are three main reasons for this. First, there is a lack of incentives. Neither journal editors nor regulatory authorities require such analyses in the reporting of the results of clinical trials with missing data. Second, there is what has been called the "knowing-doing" gap [32]. Third, there has been a lack of software tools. Recommendation 18 of the NRC report recognizes "development of software that supports coherent missing data analyses [as] a high priority".

The purpose of the grant was to bridge some of these gaps by:

1. creating methods for global sensitivity analysis of clinical trials with monotone and non-monotone missing data

2. developing free, open source and reproducible software in SAS and R to implement the methods, and

3. demonstrating the methods and software using real clinical trial data,

Given dual funding from the FDA, the focus of the funding from PCORI was on studies with patient-centered outcomes and on dissemination and translation to the PCOR community.

# D Patient/Stakeholder Engagement

An advisory board was created for the project (see Table 1). The advisory board was designed to be interdisciplinary and to represent multiple perspectives. It included experts in biostatistical methods, software development, evidence-based medicine and patient-centered outcomes. The board included academic, industry and regulatory as well as clinical and patient perspectives.

| Name | Affiliation | Expertise |
|---|---|---|
| Eric Bass | Johns Hopkins University | Evidence-Based Medicine |
| James Carpenter | London School of Hygiene | Biostatistical Methods |
| Diane Fairclough | University of Colorado | Patient-Centered Outcomes |
| Joseph Hogan | Brown University | Biostatistical Methods |
| Rod Little | University of Michigan | Biostatistical Methods |
| Devan Mehrotra | Merck | Industry, Biostatistics |
| Cyrus Mehta | Cytel Corporation | Software Development |
| Jim Neaton | University of Minnesota | Clinical trials, Biostatistics |
| Jane Permutter | Gemini Group | Patient Advocate |
| Dennis Revicki | Evidera | Patient-Centered Outcomes |
| Andrea Rotnitzky | Di Tella & Harvard Universities | Biostatistical Methods |
| Jay Siegel | Janssen | Industry, Regulatory |
| Sean Tunis | Center for Medical Technology Policy | Evidence-Based Medicine |
| Russ Wolfinger | SAS | Software Development |
| Albert Wu | Johns Hopkins University | Patient-Centered Outcomes |

Table 1: Advisory Committee

We engaged with the advisory board on an ad-hoc basis, primarily about technical issues, access to datasets with patient centered outcomes and venues for dissemination. We also engaged regularly with FDA statisticians, principally Thomas Permutt and Gregory Levin.

Examples of how we have successfully engaged with our advisors included:

- Through Cyrus Mehta, we engaged with programmers at Cytel Corporation to over-

come challenges regarding the creation of SAS procedures.

- Through interactions with the FDA, we learned that premature withdrawal tends to be the primary informative source of missing data in regulatory clinical trials and that missing data prior to last visit tends to be a second order concern. This led us to adopt a strategy of multiply imputing missing data prior to last visit on-study under a reasonable assumption that leverages all the available data, conducting rigorous sensitivity analysis on the resulting monotonized datasets, combining results across imputed datasets and using bootstrap to construct confidence intervals.

- We engaged with Dennis Revicki to understand whether there is evidence to suggest a differential impact of a unit change in Quality of Life Enjoyment Satisfaction Questionnaire, a patient centered outcome, on the hazard of drop-out based on its location on the scale.

Examples of less successful results following engagement with our advisors included:

- We found it very difficult to obtain datasets that we could use as case studies to illustrate our methods and software. We found a general unwillingness of investigators to share data. This is consistent with the observations of [18, 30].

- Through Sean Tunis, we were able to connect with a senior official at FDA to understand how to encourage statistical reviewers at FDA to use our methods and software. The official identified the so-called "knowing-doing gap" as a possible explantation for slow adoption.

# E Methods [4]

Chapter 5 of the NRC report lays out a general framework for global sensitivity analysis. In this framework, inference about treatment effects requires two types of assumptions: (i) untestable assumptions about the distribution of outcomes among those with missing data and (ii) testable assumptions that serve to increase the efficiency of estimation. Type (i) assumptions are required to "identify" parameters of interest: identification means that one can mathematically express parameters of interest (e.g., treatment arm-specific means, treatment effects) in terms of the distribution of the observed data. In other words, if one were given the distribution of the observed data and given a type (i) assumption, then one could compute the value of the parameter of interest. In the absence of identification, one cannot learn the value of the parameter of interest based only on knowledge of the distribution of the observed data. Identification implies that the parameters of interest can, *in theory*, be estimated if the sample size is large enough.

There are an infinite number of ways of positing type (i) assumptions. It is impossible to consider all such assumptions. A reasonable way of positing these assumptions is to

(a) stratify individuals with missing outcomes based on some features, and

(b) separately for each stratum, hypothesize a connection (or link) between the distribution of the missing outcomes with the distribution of these outcomes for patients who share the same features and for whom the distribution is identified.

The connection that is posited in (b) is a type (i) assumption. The problem with this approach is that the stratum of people who share the same features will typically be very small. As a result, it is necessary to draw strength across strata by "smoothing." Smoothing is required because, *in practice*, we are not working with large enough sample sizes. Without smoothing, the data analysis will not be informative because the uncertainty (i.e., standard errors) of the parameters of interest will be too large to be of substantive use. Thus, it is necessary to impose type (ii) smoothing assumptions. Type (ii) assumptions are testable (i.e., place restrictions on the distribution of the observed data) and should be scrutinized

[4] Some material in this section as been reproduced, with permission, from [44]

via model checking.

The global sensitivity framework proceeds by parameterizing (i.e., indexing) the connections (i.e., type (i) assumptions) in (b) above via sensitivity analysis parameters. The parameterization is configured so that a specific value of the sensitivity analysis parameters (typically set to zero) corresponds to a benchmark connection that is considered reasonably plausible and sensitivity analysis parameters further from the benchmark value represent more extreme departures from the benchmark connection.

## E.1 Monotone Missing Data

### E.1.1 Notation

Let $k = 0, 1, \ldots, K$ refer in chronological order to the scheduled assessment times, with $k = 0$ corresponding to baseline. Let $Y_k$ denote the outcome scheduled to be measured at assessment $k$. Define $R_k$ to be the indicator that an individual is on-study at assessment $k$. We assume that all individuals are present at baseline. Further, we assume that individuals do not contribute any further data once they have missed a visit. Let $C = \max\{k : R_k = 1\}$ and note that $C = K$ implies that the individual must have completed the study. For any given vector $z = (z_1, z_2, \ldots, z_K)$, we define $z_k = (z_0, z_1, \ldots, z_k)$ and $\underline{z}_k = (z_{k+1}, z_{k+2}, \ldots, z_K)$. For each individual, $O = (C, Y_C)$ is drawn from some distribution $P^*$ contained in the non-parametric model $\mathscr{M}$ of distributions. The observed data consist of $n$ independent draws

$O_1, O_2, \ldots, O_n$ from $P^*$. The superscript $*$ is used to denote the true value of the quantity to which it is appended.

By factorizing the distribution of $O$ in terms of chronologically ordered conditional distributions, any distribution $P \in \mathscr{M}$ can be represented by

- $F_0(y_0) := P(Y_0 \leq y_0)$;
- $F_{k+1}(y_{k+1} \mid \overline{y}_k) := P\left(Y_{k+1} \leq y_{k+1} \mid R_{k+1} = 1, \overline{Y}_k = \overline{y}_k\right)$, $k = 0, 1, \ldots, K-1$;
- $H_{k+1}(\overline{y}_k) := P\left(R_{k+1} = 0 \mid R_k = 1, \overline{Y}_k = \overline{y}_k\right)$, $k = 0, 1, \ldots, K-1$.

The main objective is to draw inference about $\mu^* := E^*(Y_K)$, the true mean outcome at visit

$K$ in a hypothetical world in which all patients are followed to that visit.[5]

### E.1.2 Assumptions

Assumptions are required to draw inference about $\mu^*$ based on the available data. We consider a class of assumptions whereby an individual's decision to drop out in the interval between visits $k$ and $k+1$ is not only influenced by past observable outcomes but by the outcome at visit $k+1$. Towards this end, we adopt the following two assumptions introduced in [41]:

**Assumption 1:** For $k = 0, 1, \ldots, K-2$,

$$P^*\left(Y_K \leq y \mid R_{k+1} = 0, R_k = 1, \overline{Y}_{k+1} = \overline{y}_{k+1}\right) \;=\; P^*\left(Y_K \leq y \mid R_{k+1} = 1, \overline{Y}_{k+1} = \overline{y}_{k+1}\right). \tag{1}$$

This says that in the cohort of patients who (a) are on-study at assessment $k$, (b) share the same outcome history through that visit and (c) have the same outcome at assessment $k+1$, the distribution of $Y_K$ is the same for those last seen at assessment $k$ and those still on-study at $k+1$.

**Assumption 2:** For $k = 0, 1, \ldots, K-1$,

$$dG^*_{k+1}(y_{k+1} \mid \overline{y}_k) \;\propto\; \exp\{\rho_{k+1}(\overline{y}_k, y_{k+1})\} dF^*_{k+1}(y_{k+1} \mid \overline{y}_k)\;, \tag{2}$$

where $G^*_{k+1}(y_{k+1} \mid \overline{y}_k) := P^*\left(Y_{k+1} \leq y_{k+1} \mid R_{k+1} = 0, R_k = 1, \overline{Y}_k = \overline{y}_k\right)$ and $\rho_{k+1}(\overline{y}_k, y_{k+1})$ is a known, pre-specified function of $y_k$ and $y_{k+1}$.

Conditional on any given history $\overline{y}_k$, Assumption 2 relates the distribution of $Y_{k+1}$ for those patients who drop out between assessments $k$ and $k+1$ to those patients who are on study at $k+1$. The special case whereby $\rho_{k+1}$ is constant in $y_{k+1}$ for all $k$ implies that, conditional on the history $y_k$, individuals who drop out between assessments $k$ and $k+1$ have the same distribution of $Y_{k+1}$ as those on-study at $k+1$. If instead $\rho_{k+1}$ is an increasing (decreasing) function of $y_{k+1}$ for some $k$, then individuals who drop-out between assessments $k$ and $k+1$ tend to have higher (lower) values of $Y_{k+1}$ than those who are on-study at $k+1$.

[5] Our methodology is developed specifically for drawing inference about the mean of the outcome. If one is interested in estimating the mean of an ordinal outcome, then our method applies; otherwise it does not.

For specified $\rho_{k+1}$, Assumptions 1 and 2 are type (i) assumptions; they place no restriction on the distribution of the observed data. As such, $\rho_{k+1}$ is not an empirically verifiable function. For simplicity, we take $\rho_{k+1}(Y_k, Y_{k+1}) = \overline{\alpha\rho}(Y_{k+1})$, where $\rho$ is a specified function of its argument and $\alpha$ is a sensitivity analysis parameter.[6]

### E.1.3 Identifiability of target parameter

Under Assumptions 1 and 2, the parameter $\mu^*$ is identifiable. To establish identifiability, it suffices to demonstrate that $\mu^*$ can be expressed as a functional of the distribution of the observed data. The functional $\mu(P^*)$ can be equivalently expressed as

$$\int_{y_0} \cdots \int_{y_K} y_K \prod_{k=0}^{K-1} \left\{ dF^*_{k+1}(y_{k+1} \mid \overline{y}_k) \left\{ 1 - H^*_{k+1}(\overline{y}_k) \right\} + \frac{\exp\{\rho_{k+1}(\overline{y}_k, y_{k+1})\} dF^*_{k+1}(y_{k+1} \mid \overline{y}_k)}{\int \exp\{\rho_{k+1}(\overline{y}_k, u)\} dF^*_{k+1}(u \mid \overline{y}_k)} H^*_{k+1}(\overline{y}_k) \right\} dF^*_0(y_0) \ . \quad (3)$$

### E.1.4 Statistical inference

Given a fixed function $\rho_{k+1}$, [41] proposed to estimate $\mu^*$ via the plug-in principle. Specifically, they specify type (ii) smoothing assumptions in the form of parametric models for both $F^*_{k+1}$ and $H^*_{k+1}$, estimate parameters in these models by maximum likelihood, estimate $F^*_0$ nonparametrically using the empirical distribution function, and finally, estimate (3) by Monte Carlo integration using repeated draws from the resulting estimates of $F^*_{k+1}$, $H^*_{k+1}$ and $F^*_0$. Since (3) is a smooth functional of $F^*_0$ and of the finite-dimensional parameters of the models for $F^*_{k+1}$ and $H^*_{k+1}$, the resulting estimator of $\mu^*$ is $n^{1/2}$-consistent and, suitably normalized, tends in distribution to a mean-zero Gaussian random variable.

While simple to describe and easy to implement, this approach has a major drawback: the inferences it generates will be sensitive to correct specification of the parametric models

[6] Our sensitivity analysis as encoded in Assumption 2 uses a device called exponential tilting. While other global sensitivity analysis approaches that we have reviewed do not specifically use this device, they are similar in spirit. For example, [24] use an affine transformation (indexed by sensitivity parameters) to connect the distribution of the outcome at visit $k+1$ conditional on past history of outcomes and drop-out at visit $k$ to the distribution of outcome at visit $k+1$ conditional on the same past history of outcomes and on-study at visit $k+1$.

imposed on $F^*_{k+1}$ and $H^*_{k+1}$. Since the fit of these models is empirically verifiable, the plausibility of the models imposed can be scrutinized in any given application. In several instances, we have found it difficult to find models providing an adequate fit to the observed data. This is a serious problem since model misspecification will generally lead to inconsistent inference, which can translate into inappropriate and misleading scientific conclusions. To provide greater robustness, we instead adopted a more flexible modeling approach.

Instead, we assumed that the distribution of the observed data $P^*$ is contained in the submodel $\mathscr{M}_0 \subset \mathscr{M}$ of distributions that exhibit a first-order Markovian structure in the sense that $F_{k+1}(y_{k+1} \mid y_k)^{-} = F_{k+1}(y_{k+1} \mid y_k)$ and $H_{k+1}(y^{-}{}_k) = H_{k+1}(y_k)$. We then estimate $F^*_{k+1}$ and $H^*_{k+1}$ by Nadaraya-Watson kernel estimators and select the associated tuning parameters by $J$-fold cross validation. The adequacy of the Markovian assumptions and the quality of the estimators for $F^*_{k+1}$ and $H^*_{k+1}$ can be assessed using goodness-of-fit procedures. The tuning parameters are generally chosen to achieve an optimal finite-sample bias-variance trade-off for the quantity requiring smoothing – here, conditional distribution and probability mass functions. However, this trade-off may be problematic, since the resulting plug-in estimator $\mu(\widehat{P^*})$ may suffer from excessive and asymptotically nonnegligible bias due to inadequate tuning. This may prevent the plug-in estimator from having regular asymptotic behavior. In particular, the resulting estimator may have a slow rate of convergence, and common methods for constructing confidence intervals, such as the Wald and bootstrap intervals, can have poor coverage properties. Therefore, the plug-in estimator must be regularized in order to serve as an appropriate basis for drawing statistical inference.

To address this problem, we employ a one-step bias correction procedure. This procedure involves adding a bias correction term to the plug-in estimator. The bias correction term is the average of the estimated "influence function", which measures the impact of "infinitesimal" contamination of $P^*$ on (3). The resulting corrected estimator can be shown to have second-order asymptotic bias that ensures regular asymptotic behavior.

To characterize the uncertainty of our estimation procedure, we utilize bootstrapping techniques.

## E.2 Non-monotone Missing Data

Studies with non-monotone patterns are much more challenging to analyze. They are typ- ically analyzed under the "missing at random" assumption. While this assumption is con- sidered reasonable for studies with monotone missing data, [36, 28] have argued that it is not appropriate for studies with non-monotone missing data. However, positing plausible assumptions and specifying flexible models for such studies is challenging because of the potentially large number of missing data patterns (as many as $2^K - 1$ patterns). Ibrahim and Molenberghs (2009) [20] indicate that "[s]uch data present a considerable modeling chal- lenge for the statistician". At the time of the project proposal, there was only one global sensitivity analysis proposal for analyzing non-monotone missing data [31]. Unfortunately, this procedure is anchored at an implausible benchmark assumption [26].

During the project period, we were unable to develop a global sensitivity analysis procedure for handling non-monotone missing data (see Section G.1). Nonetheless, we did develop the following procedure for imputing missingness prior to last visit on study. This imputation procedure is used to create datasets with monotone missing data patterns. Then the global sensitivity analysis developed in the previous section can be applied.

### E.2.1 Notation

Let $M_k$ denote the indicator that $Y_k$ is unobserved at time $k$. We assume that $M_0 = 0$ and $M_C = 0$. By construction, $M_k = 1$ if $R_k = 0$. Let $O_k = (M_k, Y_k : M_k = 0)$. The observed data for an individual are $\overline{O_K}$. With this notation, $O_0 = Y_0$ and $C$ can be computed from $\overline{O}_K$ as $\max\{k : M_k = 0\}$.

### E.2.2 Identifiability Asssumption

To handle missing data prior to last visit on study, we adapt an untestable identifiability assumption from Robins (1997) [36]. Specifically, we assume that, for $0 < k < C$, $M_k$ is independent of $Y_k$ given $\overline{Y_{k-1}}$ and $O_k$. In words, this assumption says that, while on-study, the probability of providing outcome data at time $k$ can depend on previous outcomes (observed or not) and observed data after time $k$. Alternatively, imagine a stratum of

individuals who share the same history of outcomes prior to time $k$ and same observed data after time $k$. Now, imagine splitting the stratum into two sets: those who provide outcome data at time $k$ (stratum B) and those who do not (stratum A). This assumption says that the distribution of the outcome at time $k$ is the same for these two strata. Mathematically, we write this assumption as follows:

$$dF^*(Y_k|\underbrace{M_k=1,\overline{Y}_{k-1},\underline{O}_k}_{\text{Stratum A}}) = dF^*(Y_k|\underbrace{M_k=0,\overline{Y}_{k-1},\underline{O}_k}_{\text{Stratum B}}): \quad 0<k<C. \tag{4}$$

Using Bayes' rule, (4) can be written as follows:

$$P^*(M_k=1 \mid \overline{Y}_k,\underline{O}_k) = P^*(M_k=1 \mid \overline{Y}_{k-1},\underline{O}_k): \quad 0<k<C. \tag{5}$$

Letting $\omega_k^*(\overline{Y}_{k-1},\underline{O}_k) = P^*(M_k=1 \mid \overline{Y}_{k-1},\underline{O}_k)$, it can be shown that

$$M_k \perp Y_k \mid \omega_k^*(\overline{Y}_{k-1},\underline{O}_k): \quad 0<k<C \tag{6}$$

Under assumption (4), the joint distribution of $(C,\overline{Y}_C)$ (i.e., the monotonized) is identified by a recursive algorithm.

### E.2.3 Smoothing Assumptions

We assume fully parametric restrictions on $\omega_k^*(\overline{Y}_{k-1},\underline{O}_k)$. Specifically, we assume

$$\text{logit}\{\omega_k^*(\overline{Y}_{k-1},\underline{O}_k)\} = w_k(\overline{Y}_{k-1},\underline{O}_k;\nu_k^*); \quad k=1,\ldots,K-1 \tag{7}$$

where $w_k(\overline{Y_{k-1}}, O_k;\underline{\nu_k})$ is a specified function of its arguments and $\nu_k$ is a finite-dimensional parameter with true value $\nu_k^*$.

### E.2.4 Simultaneous Estimation/Imputation

The parameters $\nu_k^*$ $(k=1,\ldots,K-1)$ can be estimated and the intermittent missingness can be imputed using the following sequential procedure:

1. Set $k=1$.

2. Estimate $\nu_k^*$ by $\widehat{\nu}_k$ as the solution to:

$$\sum_{i=1}^{n} R_{k,i} d_k(\overline{Y}_{k-1,i}, \underline{O}_{k,i}; \nu_k) \left( M_{k,i} - \text{expit}\{w_k(\overline{Y}_{k-1,i}, \underline{O}_{k,i}; \nu_k)\} \right) = 0,$$

where $d_k(\overline{Y}_{k-1}, \underline{O}_k; \nu_k^*)$ is the derivative of $w_k(\overline{Y}_{k-1}, \underline{O}_k; \nu_k)$ with respect to $\nu_k$ evaluated at $\nu_k^*$.

3. For each individual $i$ with $R_{k,i} = 1$, compute

$$\widehat{\omega}_k(\overline{Y}_{k-1,i}, \underline{O}_{k,i}) = \text{expit}\{w_k(\overline{Y}_{k-1,i}, \underline{O}_{k,i}; \widehat{\nu}_k)\}.$$

Let $\mathcal{J}_k = \{i : R_{k,i} = 1, M_{k,i} = 0\}$ and $\mathcal{J}_k' = \{i : R_{k,i} = 1, M_{k,i} = 1\}$. For each individual $i \in \mathcal{J}_k'$, impute $Y_{k,i}$ by randomly selecting an element from the set

$$\left\{ Y_{k,l} : l \in \mathcal{J}_k, \widehat{\omega}_k(\overline{Y}_{k-1,l}, \underline{O}_{k,l}) \text{ is "near" } \widehat{\omega}_k(\overline{Y}_{k-1,i}, \underline{O}_{k,i}) \right\} \tag{8}$$

4. Set $k = k + 1$. If $k = K$ then stop. Otherwise, return to Step 2.

The imputation part of this algorithm is similar in spirit to the sequential missing data imputation strategy of Lavori, Dawson and Shera (1995) [22].

We use this algorithm to create $M$ monotone missing datasets. The monotone missing data methods discussed above can then be applied to each of these datasets. Overall point estimates can be obtained by averaging across imputed datasets. To characterize the uncertainty of our estimation procedure, we utilize bootstrapping techniques. Methodology for assessing sensitivity to Assumption (4) was not developed.

## E.3 Case Study

### E.3.1 Introduction

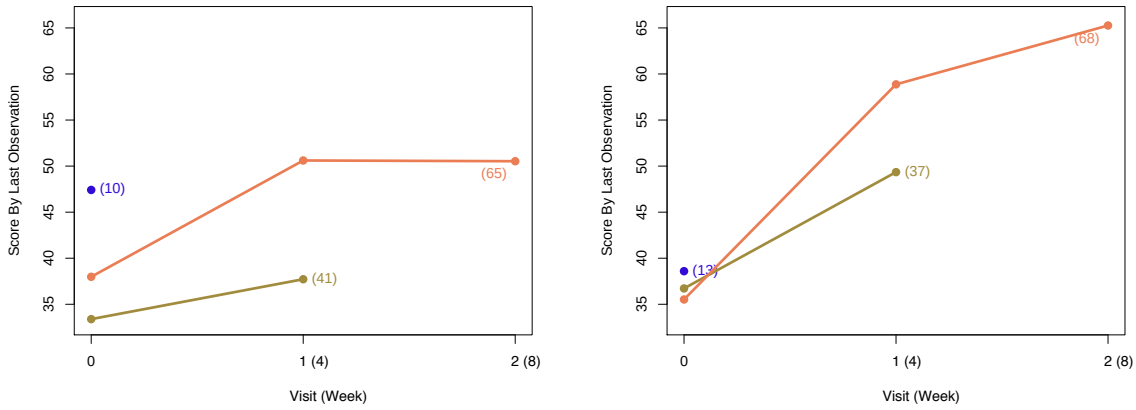
The Quetiapine Bipolar trial was a multi-center, placebo-controlled, double-blind study in which patients with bipolar disorder were randomized equally to one of three treatment arms: placebo, Quetiapine 300 mg/day or Quetiapine 600 mg/day [3]. Randomization was stratified by type of bipolar disorder: 1 or 2. A key secondary patient-reported endpoint was the short-form version of the Quality of Life Enjoyment Satisfaction Questionnaire (QLESSF,

[14]) which was scheduled to be measured at baseline (visit 0), week 4 (visit 1) and week 8 (visit 2).

We focus on the subset of 234 patients with bipolar 1 disorder who were randomized to either the placebo (n=116) or 600 mg/day (n=118) arms.[7] We seek to compare the mean QLESSF outcomes at week 8 (visit 2) between these two treatment groups, in a counterfactual world in which there are no missing outcomes. Unfortunately, this comparison is complicated because patients prematurely withdrew from the study. In this case study, all missingness is monotone (i.e., no outcomes are observed for an individual after their first missed assessment). Figure 1 displays the treatment-specific trajectories of mean QLESSF scores, stratified by the last available measurement. Notice that only 65 patients (56%) in the placebo arm and 68 patients (58%) in the 600mg/day arm had a complete set of QLESSF scores. Further, the patients with complete data tend to have higher average QLESSF scores, suggesting that a complete-case analysis could be biased.

[7]These sample sizes exclude three randomized patients - one from placebo and two from 600 mg/day Quetiapine. From each group, one patient was removed because of undue influence on the analysis. In the 600 mg/day Quetiapine arm, one patient had incomplete questionnaire data at baseline.

Figure 1: Treatment-specific (left: placebo; right: 600 mg/day Quetiapine) trajectories of mean QLESSF scores, stratified by last available measurement. Blue, brown and orange represent the trajectories of patients last seen at visits 0, 1 and 2 (weeks 0, 4 and 8), respectively. The number in parentheses at the end of each trajectory represents the number of associated patients.




### E.3.2 Analysis

The broad goal of our analytic approach is to draw inferences about treatment effects, acknowledging that, at each visit, the distribution of missing outcomes may differ from the distribution of observed outcomes. A naive analysis might assume that these distributions are the same or, at least, the same conditional on observable data. In our approach, we introduce sensitivity parameters that characterize (unknown) differences between the conditional (on observable data) distributions of missing and observed outcomes. By varying the sensitivity parameters, we can assess how sensitive the study results are to different assumptions about the distribution of missing outcomes. The final analysis then allows researchers and decision makers to have a more realistic understanding of treatment effects in the presence of missing outcome data.

Our approach has two major steps:

1. Estimate a model for the distribution of the observed data and evaluate its goodness of fit. This step involves estimating, for individuals on study at visit $k$, (a) the conditional

probability of drop-out before visit $k+1$ given the outcome measured at visit $k$ ("drop-out" model) and (b) the conditional distribution of the outcome at visit $k$ given the outcome measured at visit $k-1$ ("outcome" model). It is assumed that the dependence in (a) and (b) on previous outcomes is "smooth", i.e., previous outcome values close to one another do not have wildly different effects. Smoothing serves to increasing the precision of the ultimate inferences. The degree of smoothness is estimated from the data.

2. For each choice of treatment-specific sensitivity parameters, use the model derived in the first step to "impute" missing outcomes, estimate treatment-arm specific means and evaluate treatment effects.

In the first step, the estimated smoothing parameters for the drop-out (outcome) model are 11.54 (6.34) and 9.82 (8.05) for the placebo and 600 mg arms, respectively. In the placebo arm, the observed percentages of last being seen at visits 0 and 1 among those on-study at these visits are 8.62% and 38.68%, respectively. Estimates of last being seen at visits 0 and 1 derived from the model for the distribution of the observed data are 7.99% and 38.19%, respectively. For the 600 mg arm, the observed percentages are 11.02% and 35.24% and the model-based estimates are 11.70% and 35.08%. For both treatment arms, the maximum absolute distances between the empirical distribution of the observed outcomes and the model-based estimates of the distribution of outcomes among those on-study at visits 1 and 2 are quite small (0.013 and 0.033 for the placebo arm; 0.013 and 0.022 for 600 mg arm). These results suggest that our model for the observed data fits the observed data well and provides confidence in the second step "imputation" scheme.

To estimate the treatment-specific mean outcome at week 8 in a counterfactual world in which there are no missing data, we must impose assumptions that allow us to "impute" outcomes for those with missing data. There is an infinite number of ways of positing such assumptions. As discussed above, the treatment-arm specific strategy is as follows:

- Consider individuals last seen at visit 1 and sub-stratify these individuals by their observed QLESSF scores at visits 0 and 1. The sensitivity parameter links the unknown distribution of missing visit 2 outcomes for individuals in each sub-stratum to the

reference distribution of observed visit 2 outcomes for those individuals who complete the study and have the same visit 0 and visit 1 QLESSF scores. This latter distribution is estimated using the results of the first step.

- Consider the individuals last seen at visit 0 and sub-stratify these individuals by their observed QLESSF scores at visit 0. The sensitivity parameter links the unknown distribution of missing visit 1 outcomes for individuals in each sub-stratum to the reference distribution of observed visit 1 outcomes for those individuals who are on-study at visit 1 and have the same visit 0 score. This latter distribution is estimated using the results of the first step.

- Consider the individuals last seen at visit 0 and sub-stratify these individuals by their observed QLESSF scores at visit 0 and their counterfactual QLESSF scores at visit 1 (i.e., the visit 1 outcomes we would have observed had they not dropped out). Assume that the distribution of missing visit 1outcomes for individuals in each sub-stratum is the same as the distribution of observed visit 2 outcomes for those individuals who are on- study at visit 1 and have the same visit 0 and visit 1 scores.

The sensitivity parameter governs differences in the distributions that are being linked, with zero value indicating no difference. The zero value corresponds to the benchmark missing at random assumption. Positive (negative) values of the sensitivity parameter indicate that distribution of the missing outcomes is more skewed to higher (lower) values than the reference distribution. The skewness increases with the magnitude of the sensitivity analysis parameter.

We assume that the sensitivity parameter is the same across all the sub-strata; otherwise, reporting the results of the sensitivity analysis would be impossible. We vary the sensitivity parameter over a wide range of values (representing large departures from missing at random) to see how inferences vary.

Under missing at random, the estimated means of the outcome at visit 2 (in a counterfactual world without missing data) are 46.45 (95% CI: 42.35,50.54) and 62.87 (95% CI: 58.60,67.14) for the placebo and 600 mg arms, respectively. The estimated difference between 600 mg and placebo is 16.42 (95% 10.34, 22.51), which represents a statistically significant

improvement in quality of life in favor of Quetiapine. This difference is also clinically meaningful [15].

The treatment-arm specific sensitivity parameter in this case study can be interpreted as the log hazard ratio of drop-out between visits $k$ and $k+1$ for patients who differ by one unit in QLESSF at visit $k+1$, after controlling for the observed outcomes through visit $k$. In each treatment group, we ranged the sensitivity parameter, which we denote by $\alpha$, over a very wide range (-10 to 10). Figure 2 presents treatment-specific estimates (along with 95% confidence intervals) of the means of the outcome at visit 2 (in a counterfactual world without missing data) as a function of the sensitivity parameter.

To help interpret the sensitivity analysis parameter, Figure 3 displays treatment-specific differences between the estimated mean QLESSF at visit 2 among non-completers (back-calculated from the estimated counterfactual mean, estimated mean among completers and proportion of completers) and the estimated mean among completers, as a function of the sensitivity parameter. For example, when the sensitivity parameter equals -10, non-completers are estimated to have more than 20 points lower quality of life than completers; this holds for both treatment arms. In contrast, when the sensitivity parameter equals +10, non- completers are estimated to have 6 and 11 points higher quality of life than completers in the placebo and Quetiapine arms, respectively. Researchers can judge the plausibility of the sensitivity parameter by whether these differences seem clinically plausible. In this setting, it may be considered unreasonable that completers are worse off in terms of quality of life than non-completers, in which case the sensitivity parameter should be restricted to be less than 6 in the placebo arm and less than 3 in the Quetiapine arm.[8]

Figure 4 displays a contour plot of the estimated differences between the mean QLESSF at visit 2 (in a counterfactual world without missing data) for Quetiapine vs. placebo for various treatment-specific combinations of the sensitivity parameters. The point (0,0) corresponds to the missing at random assumption in both treatment arms. The figure shows that the differences are statistically significant (represented by dots) in favor of Quetiapine

[8] In planning a clinical trial, investigators can pre-specify a clinically plausible range for the difference in outcome between completers and non-completers; this range can then be translated into a range for the sensitivity parameters.

at almost all combinations of the sensitivity parameters. Only when the sensitivity analysis are highly differential (e.g., 8 in placebo and -8 in Quetiapine) are the differences no longer statistically significant. This figure shows that conclusions are highly robust to deviations from missing at random.

Figure 2: Treatment-specific (left: placebo; right: 600 mg/day Quentiapine) estimates (along with 95% pointwise confidence intervals) of $\mu^*$ as a function of $\alpha$.

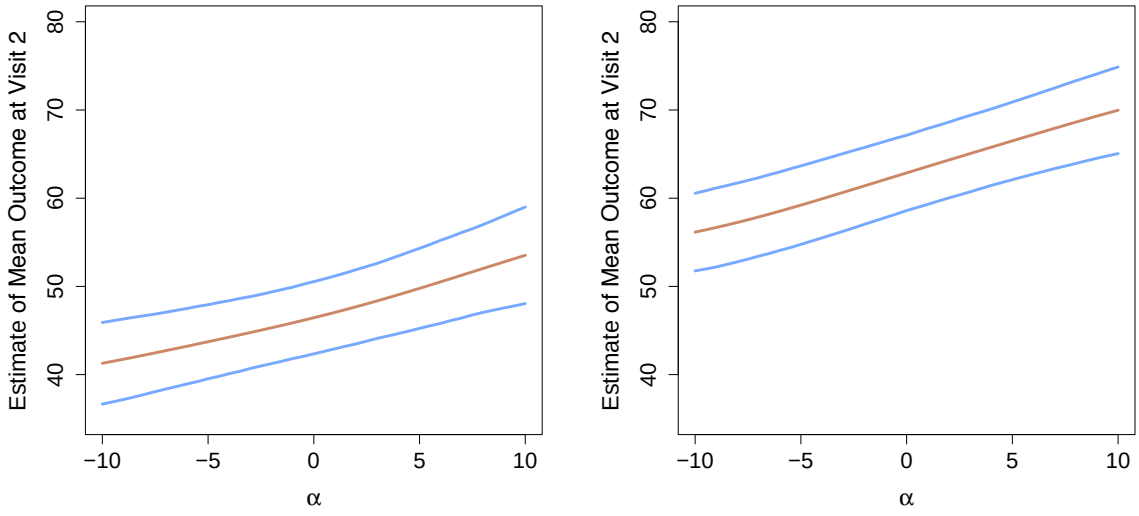


Figure 3: Treatment-specific differences between the estimated mean QLESSF at visit 2 among non-completers and the estimated mean among completers, as a function of $\alpha$.

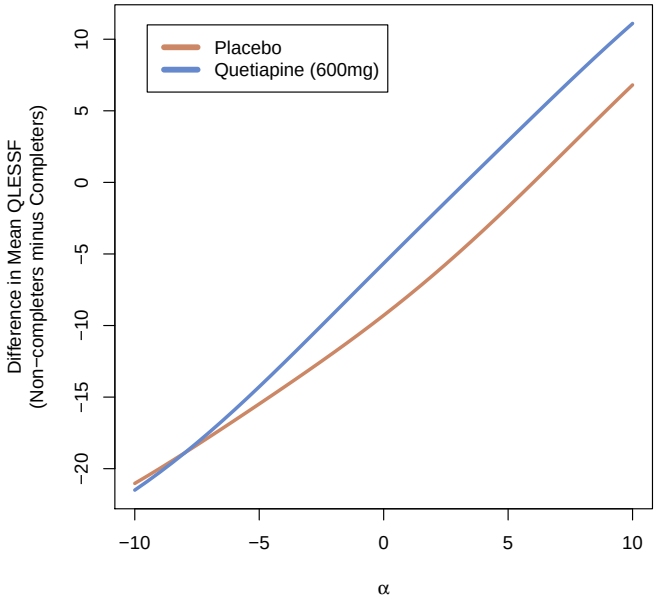
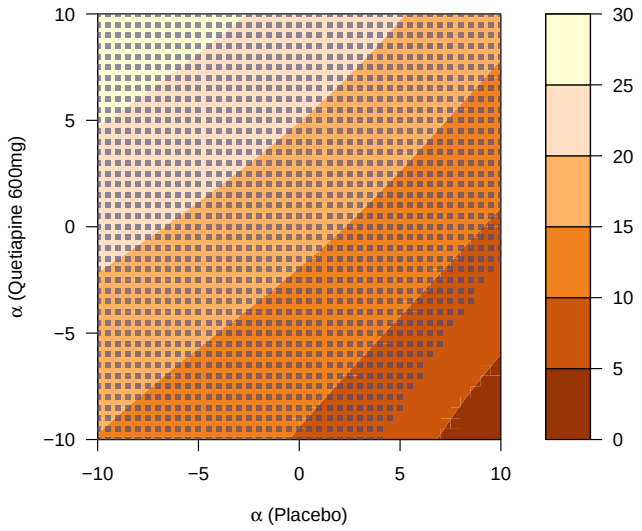



Figure 4: Contour plot of the estimated differences between mean QLESSF at visit 2 for Quentiapine vs. placebo for various treatment-specific combinations of the sensitivity analysis parameters. The point (0,0) corresponds to the MAR assumption in both treatment arms.



# F Results

The statistical methods described in Section E were implemented in the software package `SAMON`. R and SAS versions of `SAMON` are available at `www.missingdatamatters.org`. As of August, 2017, 325 individuals have registered to either use the software or receive project updates.

## F.1 Manuscripts

A technical manuscript (see Appendix A) describing the statistical methods described in Section E.1 has been published in *Biometrics*. The citation for the manuscript is:

- Scharfstein DO, McDermott A, Diaz I, Carone M, Lunardon N and Turkoz I (2018): "Global Sensitivity Analysis of Repeated Measures Studies with Informative Drop-out: A Semi-Parametric Perspective". *Biometrics*. [40]

A second manuscript (see Appendix B) that provides a more accessible explication of the statistical methods described in Section E.1 has been accepted for publication in *Statistical Methods in Medical Research*. This manuscript also provides a review of missing rates in clinical trials with patient-centered outcomes that were reported in five leading medical journals over the past 8 years. The citation for the manuscript is:

- Scharfstein DO and McDermott A (2018): "Global Sensitivity Analysis of Clinical Tri-als with Missing Patient Reported Outcomes". *Statistical Methods in Medical Research*. [44]

A book that discusses the entirety of the methods described in Section E as well as the software is currently in preparation. We plan to publish it on Leanpub, which will allow us to dynamically update the material as the methods and software continue to develop.

## F.2 Additional Manuscripts

During the course of project, we published additional manuscripts about the analysis of randomized trials with missing data.

The first manuscript relates to the trials in which each enrolled subject is expected to undergo a fixed sequence of "pass/fail" tests, one or more test results may be missing, and interest is focused on estimating the distribution of the earliest test at which a subject "passes" ("fails") that and all subsequent tests. This manuscript was motivated by tuberculosis trials. It was published in the *Annals of Applied Statistics*. The citation for the manuscript is:

- Scharfstein DO, Rotnitzky A, Abraham M, McDermott A, Chaisson R and Geiter L (2015): "On the Analysis of Tuberculosis Studies with Intermittent Missing Sputum Data". *Annals of Applied Statistics*. [42]

The second manuscript discusses how to analyze trials in which (1) patients are at high risk of death, (2) functional outcomes are scheduled to be measured on patients who survive to fixed points in time after randomization and (3) there are missing functional outcome data among survivors. This manuscript was motivated by a trial of treatment for late-stage cancer. It was published in *Biometrics*. The citation for the manuscript is:

- Wang CG, Scharfstein DO, Colantuoni E, Girard T and Yan Y (2016): "Inference in Randomized Trials with Death and Missingness". *Biometrics*. [48]

An R package called `idem` has been developed, a translational manuscript has been published in the *British Medicial Journal* and a manuscript that describes the methods and software has been revised for the *Journal of Statistical Software*. The citation for the BMJ manuscript is:

- Colantuoni EA, Scharfstein DO, Wang CG, Hashem MD, Leroux A, Needham DM, Girard TD (2018): "Statistical Methods to Compare Functional Outcomes in RCTs with High Mortality" *British Medical Journal*. [8]

Finally, we wrote a letter to the editor of *Journal of General Internal Medicine* about the challenge of missing data in randomized trials in which outcomes are scheduled to be collected from electronic health records. The citation for the letter is:

- Kharrazi A, Wang CG and Scharfstein DO (2015): "Prospective HER-Based Clinical Trials: The Challenge of Missing Data". *Journal of General Internal Medicine*. [21]

## F.3 Presentations

We gave ten oral presentations:

1. McDermott: Global Sensitivity Analysis of Repeated Measures Studies with Informative Dropout: A Semi- Parametric Approach, Joint Statistical Meetings of American Statistical Association, 8/2014.

2. Scharfstein: Global Sensitivity Analysis of Repeated Measures Studies with Informative Dropout: A Semi- Parametric Approach, Joint Statistical Meetings of American Statistical Association, University of Rochester, 9/2014.

3. Wang: Inference in Randomized Trials with Death and Missingness, FDA-Industry Workshop, 9/2014.

4. Li: Standards in the Prevention and Handling of Missing Data for Patient-Centered Outcomes Research, Johns Hopkins University, 12/2014.

5. Scharfstein: Global Sensitivity Analysis of Randomized Trials with Missing Data: A Frequentist Perspective. FDA - Center for Tobacco Products, 11/2015.

6. Scharfstein: Missing Data and Sensitivity Analyses in Randomized Trials, GlaxoSmithKline, 11/2015.

7. Scharfstein: Global Sensitivity Analysis of Randomized Trials with Missing Data: From the Software Development Trenches, National Institute of Statistical Sciences. 11/2015.

8. Scharfstein: Inference in Randomized Trials with Death and Missingness, Brown University, 4/2016.

9. Scharfstein: Analysis of Randomized Trials with Missing Data, Novartis, 12/2016.

10. Scharfstein: Global Sensitivity Analysis of Randomized Trials with Missing Data, Evidera, 3/2017.

Seven Short Courses:

1. Scharfstein: Global Sensitivity Analysis of Randomized Trials with Missing Data: Recent Advances, Deming Conference, 12/2014.

2. Scharfstein, McDermott, Wang: Analysis of Randomized Trials with Missing Data, Johns Hopkins University, 1/2015.

3. Scharfstein: Global Sensitivity Analysis of Randomized Trials with Missing Data, Society for Clinical Trials, 5/2015.

4. Scharfstein, McDermott, Wang: Analysis of Randomized Trials with Missing Data, FDA, 11/2015.

5. Scharfstein, McDermott, Wang: Analysis of Randomized Trials with Missing Data, Johns Hopkins University, 6/2016.

6. Scharfstein: Analysis of Randomized Trials with Missing Data, University of Washington, 7/2016.

7. Scharfstein, McDermott: Global Sensitivity Analysis of Randomized Trials with Missing Data, FDA, 5/2017.

Three Webinars:

1. Scharfstein: Analysis of Randomized Trials with Missing Data, American Statistical Association, 5/2016.

2. Scharfstein: Analysis of Randomized Trials with Missing Data, American Statistical Association, 9/2016.

3. Wang: Inference in Randomized Trials with Death and Missingness with Software Demonstration, American Statistical Association - New Jersey Chapter, 10/2016.

One Poster:

1. Scharfstein: Global Sensitivity Analysis of Randomized Trials with Missing Data, FDA ORSI Symposium, 4/2015.

One On-line Lecture:

1. Scharfstein, Li: Analysis of Prospective Studies with Missing Data, Johns Hopkins University, 7/2016.

## F.4 Case Studies

We used our primary methods and software to re-analyze three clinical trials:

A randomized trial designed to evaluate the efficacy and safety of once-monthly, injectable paliperidone palmitate, as monotherapy or as an adjunct to pre-study mood stabilizers or antidepressants, relative to placebo in delaying the time to relapse in patients with schizoaffective disorder. The primary outcome was patient function as measured by the Personal and Social Performance scale.

A randomized trial designed to evaluate the efficacy of different doses of Quetiapine on treating patients with bipolar disorder. A key outcome was quality of life as measured by the Quality of Life Enjoyment Satisfaction Questionnaire. See Section E.3 above.

Randomized trials designed to evaluate the efficacy of different doses of topiramate in reducing pain in patients with diabetic peripheral polyneuropathy. The primary outcome was patient reported pain, measured measured on a 100-mm Visual Analog Scale.

We also analyzed the following two clinical trials using our methods and software for analyzing trials with death and missingness:

A randomized trial to evaluate the efficacy of treatment for late-stage cancer patients. Primary outcomes included time to death and lean body mass among survivors.

A randomized trial to evaluate the efficacy of a breathing/awakening protocol on ICU patients. Primary outcomes included time to death and cognition among survivors.

# G Discussion

## G.1 Methods

In the original grant application, we proposed to specify type (ii) smoothing assumptions in the form of parametric models for both $F^*_{k+1}$ and $H^*_{k+1}$ (see Section E.1). After the contract was awarded, we discovered, in a couple of case studies, that we were unable to posit parametric models that provided adequate fits to the observed data. This led us to develop methods that rely on more flexible models using a first-order Markovian assumptions and non-parametric smoothing, [9] [10] This required major methodological development since the standard plug-in estimator was no longer guaranteed to have adequate large sample properties.

Since the first-order Markovian assumptions impose restrictions on the distribution of the observed data, they should be subjected to goodness-of-fit tests. This can be achieved by (1) simulating a large synthetic dataset under the estimated model for distribution of the observed data and (2) comparing summary statistics from synthetic dataset to summary statistics from the observed dataset. If the first-order Markovian assumptions fail, there should be great disparity between the summary statistics. In this case, our approach, as currently configured, will not be suitable. Further research is needed to relax the first-order

[9] The first-order Markovian assumptions allows our method to handle studies with a large number of scheduled visits. In [40], we used our methodology to analyze a randomized trials with 16 scheduled visits with 160-170 individuals per treatment arm.

[10] A common way of analyzing longitudinal data is via fully parametric mixed effects models for the outcomes. [12] These models do not impose first-order Markovian restrictions on the joint distribution of the outcomes (i.e., the conditional distribution of the outcome at time $k$ depends on the entire history of the outcomes prior to time $k$ in a parametric fashion, not just the outcome at time $k-1$. Such models are typically analyzed under the (untestable) missing at random assumption. Under missing at random, testable restrictions are imposed on the distribution of the observed data and these should be tested using goodness of fit procedures (see, for example, [7]). Some researchers have proposed fitting parametric joint models for the outcome and missingness processes, see for example [13, 10]. Unless one has strong reasons to believe the parametric modeling assumptions, one should not use these to test the validity of the missing at random assumption; rather, sensitivity analysis (as proposed in this project) or construction of bounds is recommended. [34]

Markovian in a flexible fashion, say using single index models. [6].

Using our new methods, we found, in realistic simulation studies, that standard Wald-type confidence intervals did not provide adequate coverage. This led us to explore resampling-based techniques for constructing confidence intervals. Ultimately, we discovered that confidence intervals constructed using a combination of jackknife standard errors coupled with symmetric parametric bootstrap provided reasonable coverage.

We have found that our new procedure can be sensitive to outliers. That is, there can be observations in a given dataset that can have excessive influence on the results. To date, we have not yet found a data-adaptive solution to this problem. We plan to explore this issue in the future.

In the original grant application, we planned to develop separate sensitivity analysis procedures for studies with non-monotone missing data. These procedures were also to be based on parametric models for the distribution of the observed data. Given the problem discussed above and the fact that, in many studies, missingness prior to last visit on study is a second order issue, we decided to adopt the imputation strategy discussed in Section E.2.

## G.2 Software

We have developed R and SAS software tools that implement our methods. Creating a SAS procedure was difficult. SAS requires the use of the SAS/TOOLKIT software for developing SAS procedures. The resources for using this software are extremely limited. SAS provides a "Usage and Reference" manual for the SAS/TOOLKIT which was last updated in 1991. The examples in the manual were outdated with regards to how to set up the computing environment for compiling SAS procedures. Moreover, there were no helpful online discussions about developing SAS procedures using the SAS/TOOLKIT. In addition, developing SAS procedures requires understanding very specific SAS procedure grammar. This grammar is different from common SAS analysis grammar. The learning curve for understanding the grammar was quite steep, especially in light of a lack of helpful resources. Ultimately, a programmer from Cytel was able to provide us with useful information.

To date, we have received very little feedback on our software. The FDA hired a summer intern to review our software package. We addressed the feedback in version 4 of the software. One of the aims of our proposal was to develop methods, software and case studies

based on user feedback. No substantial feedback has been received that has warranted such developments.

## G.3 Case Studies

We developed three case studies associated with the primary methods and software developed as part of this contract. Despite concerted efforts to obtain more datasets, we have found a general unwillingness in the scientific community to share data.

## G.4 Manuscripts

We were successful to publishing two manuscripts that describe the theory and methods of our global sensitivity analysis procedure for monotone missing data. Since we changed our strategy for intermittent missing data, we do not view the associated method as a stand alone technical manuscript. Rather, we plan to discuss our strategy as a separate chapter in the book we plan to publish. While we originally planned to publish this book in a traditional format via Cambridge University Press, we now plan to publish the book via `Leanpub`, as this will allow us to dynamically update it as our methods and software change. The is currently more than two-thirds complete and our plan is to submit for publication in 2019.

## G.5 Dissemination

To date, uptake of our methods and software has been slower than anticipated. Despite publicity efforts and lack of competitors, we have not been able to generate uptake of our technology. We attribute this to (1) lack of incentives and (2) "knowing-doing" gap (per our discussion with a senior FDA official) and (3) inadequate knowledge translational by statis- tical methodologists to both principal investigators and their statistical collaborators [33]. By incentives, we mean *requirements* by regulatory authorities (e.g., FDA, EMEA), fund- ing agencies (e.g., NIH, PCORI) and journal editors to rigorously evaluate the robustness of study findings to missing data assumptions. Historically, such requirements have been imperfectly effective. For example, many journals require authors to follow the CONSORT

statement [2] when preparing manuscripts for publication. There has been some empirical evidence that this has led to a general improvement of the reporting of randomized clinical trials [46]. The CONSORT statement discusses how subgroup analyses should be appropriately reported. However, Gabler *et al.* (2016) [17] has not found that there has been no improvement in the reporting of subgroup analyses. Another example is the FDA requirement of timely reporting of results of applicable clinical trials to `clinicaltrials.gov`. Anderson *et al.* (2015) [1] found that most applicable trials did not adhere to the requirement, with industry-funded trials having a better compliance rate than government/academic-funded trials.

While the National Research Council [25], regulatory agencies, PCORI and a *New England Journal of Medicine* article [27] have argued for the importance of conducting rigorous sensitivity analysis of trials with missing data, such analyses are rarely conducted. As discussed with our advisory board, regulatory and funding organizations and journals will need to require rigorous sensitivity analysis in order for such analyses to be more routinely conducted. Many study investogators know about our technology; unless required, they will be hesitant to stress-test their studies for fear that they might not be robust. As a result, use of technology is a "hard sell."

### G.6 Future Research

In the future, we plan to work on methods for robustifying our sensitivity analysis procedure so that is less sensitive to outliers. Specifically, we plan to investigate the use of approaches developed in the robust statistics literature [19, 5]. In addition, we plan to develop methods and software for relaxing the first-order Markovian assumption and assessing the sensitivity of inference to non-monotone missing data assumptions.

## H Conclusions

In this three year contract, we developed and disseminated methods and software for conducting global sensitivity analysis of randomized trials in which outcomes, scheduled to be collected at fixed points in time after randomization, are subject to missingness. While up-

take of our technology has been slower than anticipated, we believe that our unique tool is well positioned to meet the needs of study investigators once regulators, funders and journal editors begin to follow recommendations that call for mandatory reporting of sensitivity analyses.

Additional work is needed to improve our methodology including a robust, data adaptive way of handling outliers and incorporating time-independent (i.e., baseline) and time-dependent auxiliary variables. In addition, methods and software for conducting sensitivity analysis of randomized trials with highly non-monotone missing data and off-schedule visits need to be developed.

# I References


[1] Anderson, M. L., Chiswell, K., Peterson, E. D., Tasneem, A., Topping, J., and Califf, R. M. Compliance with results reporting at clinicaltrials. gov. *New England Journal of Medicine 372*, 11 (2015), 1031–1039.

[2] Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D., et al. Improving the quality of reporting of randomized controlled trials: the consort statement. *Jama 276*, 8 (1996), 637–639.

[3] Calabrese, J. R., Keck Jr, P. E., Macfadden, W., Minkwitz, M., Ketter, T. A., Weisler, R. H., Cutler, A. J., McCoy, R., Wilson, E., Mullen, J., et al. A randomized, double-blind, placebo-controlled trial of quetiapine in the treatment of bipolar i or ii depression. *American Journal of Psychiatry* (2005).

[4] Campbell, G., Pennello, G., and Yue, L. Missing data in the regulation of medical devices. *Journal of Biopharmaceutical Statistics 21*, 2 (2011), 180–195.

[5] Carroll, R. J. *Transformation and weighting in regression*. Routledge, 2017.

[6] Chiang, C.-T., and Huang, M.-Y. New estimation and inference procedures for a single-index conditional distribution model. *Journal of Multivariate Analysis 111* (2012), 271–285.

[7] Claeskens, G., and Hart, J. D. Goodness-of-fit tests in mixed models. *Test 18*, 2 (2009), 213–239.

[8] Colantuoni, E., Scharfstein, D. O., Wang, C., Hashem, M. D., Leroux, A., Needham, D. M., and Girard, T. D. Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *bmj 360* (2018), j5748.

[9] Copas, J., and Eguchi, S. Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B 63*, 871-895 (2001).

[10] Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M. G. A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. *Biometrical Journal 52*, 1 (2010), 111–125.

[11] Daniels, M., and Hogan, J. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. CRC Press, 2008.

[12] Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., et al. *Analysis of longitudinal data*. Oxford University Press, 2002.

[13] Diggle, P., and Kenward, M. Informative drop-out in longitudinal data analysis. *Applied Statistics 43* (1994), 49–93.

[14] Endicott, J., Nee, J., Harrison, W., and Blumenthal, R. Quality of life enjoyment and satisfaction questionnaire. *Psychopharmacol Bull 29*, 2 (1993), 321–326.

[15] Endicott, J., Paulsson, B., Gustafsson, U., Schi¨oler, H., and Hassan, M. Quetiapine monotherapy in the treatment of depressive episodes of bipolar i and ii disorder: improvements in quality of life and quality of sleep. *Journal of affective disorders 111*, 2-3 (2008), 306–319.

[16] Fielding, S., Maclennan, G., Cook, J. A., and Ramsay, C. R. A review of rcts in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials 9*, 1 (2008), 51.

[17] Gabler, N. B., Duan, N., Raneses, E., Suttner, L., Ciarametaro, M., Cooney, E., Dubois, R. W., Halpern, S. D., and Kravitz, R. L. No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. *Trials 17*, 1 (2016), 320.

[18] Goodman, S. N. Clinical trial data sharing: what do we do now? *Annals of internal medicine 162*, 4 (2015), 308–309.

[19] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. *Robust statistics: the approach based on influence functions*, vol. 196. John Wiley & Sons, 2011.

[20] Ibrahim, J. G., and Molenberghs, G. Missing data methods in longitudinal studies: a review. *Test 18*, 1 (2009), 1–43.

[21] Kharrazi, H., Wang, C., and Scharfstein, D. Prospective ehr-based clinical trials: the challenge of missing data. *Journal of General Internal Medicine 29*, 7 (2014),

[22] Lavori, P. W., Dawson, R., and Shera, D. A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine 14*, 17 (1995), 1913–1925.

[23] Li, T., Hutfless, S., Dickerson, K., Scharfstein, D., Neaton, J., Hogan, J., Little, R., Daniels, M., Roy, J., Mor, V., and Law, A. Minimal standards in the prevention and handling of missing data in observational and experimental patient centered outcomes research. *Washington, DC: Patient-Centered Outcomes Research Institute* (2012).

[24] Linero, A. R., and Daniels, M. J. A flexible bayesian approach to monotone missing data in longitudinal studies with nonignorable missingness with application to an acute schizophrenia clinical trial. *Journal of the American Statistical Association 110*, 509 (2015), 45–55.

[25] Little, R., Cohen, M., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., Hogan, J., Molenberghs, G., Murphy, S., Neaton, J., Rotnitzky, A., Scharfstein, D., Shih, W., Siegel, J., and Stern, H. *The Prevention and Treatment of Missing Data in Clinical Trials.* The National Academies Press, 2010.

[26] Little, R. J. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association 88*, 421 (1993), 125–134.

[27] Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., and Stern, H. The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med. 367*, 14 (Oct 2012), 1355–1360.

[28] Little, R. J., and Rubin, D. B. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[29] Ma, G., Toxel, A., and Heitjan, D. An index of local sensitivity to nonignorable drop-out in longitudinal modelling. *Statistics in Medicine 24* (2005), 2129–2150.

[30] Mbuagbaw, L., Foster, G., Cheng, J., and Thabane, L. Challenges to complete and useful data sharing. *Trials 18*, 1 (2017), 71.

[31] Minini, P., and Chavance, M. Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics 5*, 4 (2004), 531–544.

[32] Pfeffer, J., and Sutton, R. I. *The knowing-doing gap: How smart companies turn knowledge into action*. Harvard Business Press, 2013.

[33] Pullenayegum, E. M., Platt, R. W., Barwick, M., Feldman, B. M., Offringa, M., and Thabane, L. Knowledge translation in biostatistics: a survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods. *Statistics in medicine 35*, 6 (2016), 805–818.

[34] Rhoads, C. H. Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy 3*, 1 (2012).

[35] Robins, J., Rotnitzky, A., and Scharfstein, D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models for Epidemiology*, E. Halloran, Ed. Springer-Verlag, 2000, pp. 1–94.

[36] Robins, J. M. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine 16*, 1 (1997), 21–37.

[37] Rombach, I., Rivero-Arias, O., Gray, A. M., Jenkinson, C., and Burke, O. The current practice of handling and reporting missing outcome data in eight widely used proms in rct publications: a review of the current literature. *Quality of Life Research 25*, 7 (2016), 1613–1623.

[38] Rotnitzky, A., Robins, J., and Scharfstein, D. Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association 93* (1998), 1321–1339.

[39] Rotnitzky, A., Scharfstein, D., Su, T., and Robins, J. A sensitivity analysis methodology for randomized trials with potentially non-ignorable cause-specific censoring. *Biometrics 57* (2001), 103–113.

[40] Scharfstein, D., McDermott, A., Diaz, I., M, C., Lunardon, N., and Turkoz, I. Global sensitivity analysis for repeated measures studies with informative drop-out: A semi-parametric approach. *Biometrics 74* (2017), 207–219.

[41] Scharfstein, D., McDermott, A., Olson, W., and Wiegand, F. Global sensitivity analysis for repeated measures studies with informative drop-out. *Statistics in Biopharmaceutical Research 6* (2014), 338–348.

[42] Scharfstein, D., Rotnitzky, A., Abraham, M., McDermott, A., Chaisson, R., and Geiter, L. On the analysis of tuberculosis studies with intermittent missing sputum data. *The Annals of Applied Statistics 9*, 4 (2015), 2215–2236.

[43] Scharfstein, D., Rotnitzky, A., and Robins, J. Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion). *Journal of the American Statistical Association 94* (1999), 1096–1146.

[44] Scharfstein, D. O., and McDermott, A. Global sensitivity analysis of clinical trials with missing patient reported outcomes. *Statistical Methods in Medical Research (To Appear)* (2018).

[45] Troxel, A., Ma, G., and Heitjan, D. An index of local sensitivity to nonignorability. *Statistica Sinica 14* (2004), 1221–1237.

[46] Turner, L., Moher, D., Shamseer, L., Weeks, L., Peters, J., Plint, A., Altman, D. G., and Schulz, K. F. The influence of consort on the quality of reporting of randomised controlled trials: an updated review. *Trials 12*, S1 (2011), A47.

[47] Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics 57* (2001), 7–14.

[48] Wang, C., Scharfstein, D. O., Colantuoni, E., Girard, T. D., and Yan, Y. Inference in randomized trials with death and missingness. *Biometrics* (2016).

[49] Yan, X., Lee, S., and Li, N. Missing data handling methods in medical device clinical trials. *Journal of Biopharmaceutical Statistics 19* (2009), 1085–1098.

*Disclaimer:*

*The [views, statements, opinions] presented in this report are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors or Methodology Committee.*

*Acknowledgement:*

*Research reported in this report was [partially] funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (#ME-1303-6016) Further information available at:*
*https://www.pcori.org/research-results/2013/new-methods-and-software-determine-impact-missing-data-clinical-trials*