Biometrics 000, 000–000 DOI: 000
000 0000

Global Sensitivity Analysis for Repeated Measures Studies with Informative Drop-out: A Semi-Parametric Approach

Daniel Scharfstein[1,*], Aidan McDermott[1], Ivan Diaz[2], Marco Carone[3], Nicola Lunardon[4] and Ibrahim Turkoz[5]

[1]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, U.S.A.

[2]Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY, U.S.A.

[3]University of Washington School of Public Health, Seattle, WA, U.S.A.

[4]Ca' Foscari University of Venice, Venice, Italy

[5]Janssen Research and Development, LLC, Titusville, NJ, U.S.A.

email: dscharf@jhu.edu

Summary: In practice, both testable and untestable assumptions are generally required to draw inference about the mean outcome measured at the final scheduled visit in a repeated measures study with drop-out. Scharfstein *et al.* (2014) proposed a sensitivity analysis methodology to determine the robustness of conclusions within a class of untestable assumptions. In their approach, the untestable and testable assumptions were guaranteed to be compatible; their testable assumptions were based on a fully parametric model for the distribution of the observable data. While convenient, these parametric assumptions have proven especially restrictive in empirical research. Here, we relax their distributional assumptions and provide a more flexible, semi-parametric approach. We illustrate our proposal in the context of a randomized trial for evaluating a treatment of schizoaffective disorder.

Key words: Bootstrap; Cross-Validation; Exponential Tilting; Jackknife; Identifiability; One-Step Estimator; Plug-In Estimator; Selection Bias

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

We consider a prospective cohort study design in which outcomes are scheduled to be collected at fixed time-points after enrollment and the parameter of interest is the mean outcome at the last scheduled study visit. We are concerned with drawing inference about this target parameter in the setting where some study participants prematurely stop providing outcome data.

Identifiability of the target parameter requires untestable assumptions about the nature of the process that leads to premature withdrawal. A common benchmark assumption, introduced by Rubin (1976), is that a patient's decision to withdraw between visits k and $k+1$ depends on outcomes through visit k (i.e., past), but not outcomes after visit k (i.e., future). This assumption has been referred to as *missing at random* (MAR). A weaker version of this assumption, termed *sequential ignorability* (SI), posits that the withdrawal decision depends on outcomes through visit k, but not the outcome at the last scheduled study visit (Birmingham et al., 2003). MAR yields identification of the entire joint distribution of the outcomes, while SI only admits identification of the distribution of the outcome at the last scheduled visit. Both parametric (e.g., Schafer, 1997; Little and Rubin, 2014) and semi-parametric (e.g., van der Laan and Robins, 2003; Tsiatis, 2006) approaches have been proposed for drawing inference about the target parameter under these assumptions.

For such untestable assumptions, it is important to conduct a sensitivity analysis to evaluate the robustness of the resulting inferences (e.g., Little et al., 2010; ICH, 1998; CHMP, 2009). As reviewed by Scharfstein et al. (2014), sensitivity analyses can generally be classified as ad-hoc, local and global. Ad-hoc sensitivity analysis involves analyzing the data using a variety of methods and evaluating whether the inferences they yield are consistent with one another. Local sensitivity analysis evaluates how inferences vary in a small neighborhood of

the benchmark assumption. In contrast, global sensitivity analysis considers how inferences vary over a much larger neighborhood of the benchmark assumption.

In addition to untestable assumptions, testable restrictions are needed to combat the so-called "curse of dimensionality" (Robins et al., 1997). Scharfstein et al. (2014) developed a global sensitivity analysis approach whereby the untestable and testable assumptions were guaranteed to be compatible. Their testable assumptions were based on a fully parametric model for the distribution of the observable data. In practice, we have found it particularly challenging to posit parametric models that correspond well with the observed data, as we illustrate in Section 4 below. This has motivated the current paper, in which we relax distributional assumptions and develop a more flexible, semi-parametric extension of the Scharfstein et al. (2014) approach. The techniques of Daniels and Hogan (2008) and Linero and Daniels (2015) provide Bayesian solutions to the same problem and also ensure the compatibility of the untestable and testable assumptions. However, the scalability of their approach to settings with a large number of post-baseline assessments has yet to be demonstrated.

In Section 2, we introduce the data structure and define the target parameter of interest. We also review the identification assumptions of Scharfstein *et al.* (2014). In Section 3, we present our inferential approach. In Section 4, we present results from the reanalysis of a clinical trial in which there was substantial premature withdrawal. In Section 5, we describe the results of a simulation study. We provide concluding remarks in Section 6.

2. Data structure, target parameter, assumptions and identifiability

2.1 *Data structure and target parameter*

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. Furthermore, we assume that individuals do not contribute any further data once they have missed a visit. This pattern is often referred to as monotone drop-out. 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 $\overline{z}_k = (z_0, z_1, \ldots, z_k)$ and $\underline{z}_k = (z_{k+1}, z_{k+2}, \ldots, z_K)$. For each individual, $O = (C, \overline{Y}_C)$ 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^*. Throughout, the superscript $*$ will be 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

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

Our 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.

2.2 *Assumptions*

Assumptions are required to draw inference about μ^* 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 Scharfstein et al. (2014):

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

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

This says that in the cohort of patients who (1) are on-study at assessment k, (2) share the

same outcome history through that visit and (3) 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 $k = 0, 1, \ldots, K-1$,

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

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

Conditional on any given history $\overline{y}_k$, this assumption 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 ρ_{k+1} is constant in y_{k+1} for all k implies that, conditional on the history $\overline{y}_k$, 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 ρ_{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 $k+1$.

Setting $\ell^*_{k+1}(\overline{y}_k) := \text{logit}\left\{H^*_{k+1}(\overline{y}_k)\right\} - \log\left\{\int \exp\{\rho_{k+1}(\overline{y}_k, u)\} dF^*_{k+1}(u \mid \overline{y}_k)\right\}$, it can be shown that Assumptions 1 and 2 jointly imply that

$$\text{logit}\left\{P^*\left(R_{k+1} = 0 \,\middle|\, R_k = 1, \overline{Y}_{k+1} = \overline{y}_{k+1}, Y_K = y_K\right)\right\} = \ell^*_{k+1}(\overline{y}_k) + \rho_{k+1}(\overline{y}_k, y_{k+1}) \;.$$

We note that since H^*_{k+1} and F^*_{k+1} are identified from the distribution of the observed data, so is $\ell^*_{k+1}(\overline{y}_k)$. Furthermore, we observe that ρ_{k+1} quantifies the influence of Y_{k+1} on the risk of dropping out between assessments k and $k+1$, after controlling for the past history $\overline{y}_k$. In particular, Y_K is seen to not additionally influence this risk. When ρ_{k+1} does not depend on y_{k+1}, we obtain an assumption weaker than MAR but stronger than SI – we refer to it as SI-1. Under SI-1, the decision to withdraw between visits k and $k+1$ depends on outcomes through visit k but not on the outcomes at visits $k+1$ and K. For specified ρ_{k+1},

Assumptions 1 and 2 place no restriction on the distribution of the observed data. As such, ρ_{k+1} is not an empirically verifiable function.

Assumptions 1 and 2 allow the existence of unmeasured common causes of $Y_0, Y_1, \ldots, Y_K$, but does not allow these causes to directly impact, for patients on study at visit k, the decision to drop out before visit $k+1$. This is no different than under MAR or SI. To allow for a direct impact, one could utilize the sensitivity analysis model of Rotnitzky et al. (1998) which specifies

$$\text{logit}\left\{P^*\left(R_{k+1}=0 \,\middle|\, R_k=1, \overline{Y}_k=\overline{y}_k, Y_K=y_K\right)\right\} = h^*_{k+1}(\overline{y}_k) + q_{k+1}(\overline{y}_k, y_K)\ ,$$

where $h^*_{k+1}(\overline{y}_k) := \text{logit}\left\{H^*_{k+1}(\overline{y}_k)\right\} - \log\left\{\int \exp\{\rho_{k+1}(\overline{y}_k,u)\}dF^*_{K,k}(u \mid R_k=1,\overline{y}_k)\right\}$ and $F^*_{K,k}(u \mid R_k=1,\overline{y}_k) := P^*(Y_K \leqslant u | R_k=1,\overline{Y}_k=\overline{y}_k)$. Here, $q_{k+1}(\overline{y}_k, y_K)$ quantifies the influence of the outcome scheduled to be measured at the end of the study on the conditional hazard of last being seen at visit k given the observable past $\overline{y}_k$. The key disadvantage of this model is that we have found that it is challenging for scientific experts to articulate how a distal endpoint affects a more proximal event (i.e., drop-out).

2.3 *Identifiability of target parameter*

Under Assumptions 1 and 2 with given ρ_{k+1}, the parameter μ^* is identifiable. To establish identifiability, it suffices to demonstrate that μ^* can be expressed as a functional of the distribution of the observed data. In the current setting, this follows immediately by noting, through repeated applications of the law of iterated expectations, that

$$\mu^* = \mu(P^*) = E^*\left(\frac{R_K Y_K}{\prod_{k=0}^{K-1}[1+\exp\{\ell^*_{k+1}(\overline{Y}_k)+\rho_{k+1}(\overline{Y}_k, Y_{k+1})\}]^{-1}}\right)$$

The functional $\mu(P^*)$ can be equivalently expressed as

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

3. Statistical inference

3.1 *Plug-in estimator*

Given a fixed function ρ_{k+1}, Scharfstein et al. (2014) proposed to estimate μ^* via the plug-in principle. Specifically, they specify 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 (1) by Monte Carlo integration using repeated draws from the resulting estimates of F^*_{k+1}, H^*_{k+1} and F^*_0. Since (1) 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 μ^* 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 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 adopt a more flexible modeling approach.

As noted above, the distribution P^* can be represented in terms of $\{(F^*_{k+1}, H^*_{k+1}) : k = 0, 1, \ldots, K-1\}$. Suppose that 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 $F_{k+1}(y_{k+1} \mid \overline{y}_k) = F_{k+1}(y_{k+1} \mid y_k)$ and $H_{k+1}(\bar{y}_k) = H_{k+1}(y_k)$. We can then estimate F^*_{k+1} and H^*_{k+1} by Nadaraya-Watson kernel estimators of the form:

$$\widehat{F}_{k+1,\lambda_F}(y_{k+1} \mid y_k) := \frac{\sum_{i=1}^n R_{k+1,i} I(Y_{k+1,i} \leqslant y_{k+1}) \phi_{\lambda_F}(Y_{k,i} - y_k)}{\sum_{i=1}^n R_{k+1,i} \phi_{\lambda_F}(Y_{k,i} - y_k)} \quad \text{and} \tag{2}$$

$$\widehat{H}_{k+1,\lambda_H}(y_k) := \frac{\sum_{i=1}^n R_{k,i}(1 - R_{k+1,i}) \phi_{\lambda_F}(Y_{k,i} - y_k)}{\sum_{i=1}^n R_{k,i} \phi_{\lambda_F}(Y_{k,i} - y_k)}\,, \tag{3}$$

where ϕ is a symmetric probability density function, ϕ_λ refers to the rescaled density $y \mapsto \phi(y/\lambda)/\lambda$, and (λ_F, λ_H) is a vector of tuning parameters. In practice, the values of these tuning parameters need to be carefully chosen to ensure the resulting estimators of F^*_{k+1} and H^*_{k+1} perform well. As discussed next, we select the tuning parameters via J-fold cross validation.

Writing $F := (F_1, F_2, \ldots, F_K)$ and $H := (H_1, H_2, \ldots, H_K)$, and denoting a typical realization of the prototypical data unit as $o = (c, \overline{y}_c)$, we may define the loss functions

$$\begin{aligned} L_F(F; F^\circ)(o) &:= \sum_{k=0}^{K-1} r_{k+1} \int \{I(y_{k+1} \leqslant u) - F_{k+1}(u \mid y_k)\}^2 \, dF^\circ_{k+1}(u) \, , \\ L_H(H; H^\circ)(o) &:= \sum_{k=0}^{K-1} r_k \left[r_{k+1} - \{1 - H_{k+1}(y_k)\}\right]^2 H^\circ_{k+1} \end{aligned}$$

with $F^\circ := (F^\circ_1, F^\circ_2, \ldots, F^\circ_K)$ and $H^\circ := (H^\circ_1, H^\circ_2, \ldots, H^\circ_K)$ defined by $F^\circ_{k+1}(u) := P(Y_{k+1} \leqslant u \mid R_{k+1} = 1)$ and $H^\circ_{k+1} := P(R_{k+1} = 0 \mid R_k = 1)$. Here, F° and H° represent collections of distributions and probabilities that can be estimated nonparametrically without the need for smoothing. It can be shown that the true risk mappings $F \mapsto E^*\{L_F(F; F^{\circ *})(O)\}$ and $H \mapsto E^*\{L_H(H; H^{\circ *})(O)\}$ are minimized at $F = F^*$ and $H = H^*$, where $F^{\circ *}$ and $H^{\circ *}$ denote the true value of F° and H°, respectively. Given a random partition of the dataset into J validation samples $\{V_1, V_2, \ldots, V_J\}$ with sample sizes $n_1, n_2 \ldots, n_J$, taken to be approximately equal, the oracle selectors for λ_F and λ_H are (van der Vaart et al., 2006)

$$\widetilde{\lambda}_F := \underset{\lambda_F}{\operatorname{argmin}} \frac{1}{J} \sum_{j=1}^{J} E^*\{L_F(\widehat{F}^{(j)}_{\lambda_F}; \widehat{F}^\circ)(O)\} \;\text{ and }\; \widetilde{\lambda}_H := \underset{\lambda_H}{\operatorname{argmin}} \frac{1}{J} \sum_{j=1}^{J} E^*\{L_H(\widehat{H}^{(j)}_{\lambda_H}; \widehat{H}^\circ)(O)\}.$$

Here, $\widehat{F}^{(j)}_{k+1,\lambda_F}$ and $\widehat{H}^{(j)}_{k+1,\lambda_H}$ are obtained by computing (2) and (3), respectively, on the dataset excluding individuals in V_j. The estimates of nuisance parameter estimators $\widehat{F}^\circ_{k+1}$ and $\widehat{H}^\circ_{k+1}$ are given by the empirical distribution of the observed values of Y_{k+1} within the subset of individuals with $R_{k+1} = 1$ and by the empirical proportion of individuals with $R_{k+1} = 0$ among those with $R_k = 1$, respectively. The quantities $\widetilde{\lambda}_F$ and $\widetilde{\lambda}_H$ cannot be computed in

practice since P^* is unknown. Empirical tuning parameter selectors are given by

$$\widehat{\lambda}_F := \underset{\lambda_F}{\operatorname{argmin}}\, \widehat{\mathcal{R}}_F(\lambda_F) \;\text{ and }\; \widehat{\lambda}_H := \underset{\lambda_H}{\operatorname{argmin}}\, \widehat{\mathcal{R}}_H(\lambda_H),$$

where

$$\begin{aligned}\widehat{\mathcal{R}}_F(\lambda_F) &:= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i\in V_j} L_F(\widehat{F}^{(j)}_{\lambda_F};\widehat{F}^{\circ})(O_i)\\ &= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i\in V_j}\sum_{k=0}^{K-1} R_{k+1,i}\left(\frac{\sum_\ell R_{k+1,\ell}\{I(Y_{k+1,i}\leqslant Y_{k+1,l}) - \widehat{F}^{(j)}_{k+1,\lambda_F}(Y_{k+1,l}\mid Y_{k,i})\}^2}{\sum_\ell R_{k+1,\ell}}\right)\end{aligned}$$

and

$$\begin{aligned}\widehat{\mathcal{R}}_H(\lambda_H) &:= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i\in V_j} L_H(\widehat{H}^{(j)}_{\lambda_H};\widehat{H}^{\circ})(O_i)\\ &= \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i\in V_j}\sum_{k=0}^{K-1}\frac{R_{k,i}[R_{k+1,i}-\{1-\widehat{H}^{(j)}_{k+1,\lambda_H}(Y_{k,i})\}]^2\sum_\ell R_{k,\ell}(1-R_{k+1,\ell})}{\sum_\ell R_{k,\ell}}\,.\end{aligned}$$

The naive substitution estimator of μ^* is $\mu(\widehat{P})$, where $\widehat{P}$ is determined by (2) and (3) computed with tuning parameters $(\widehat{\lambda}_F, \widehat{\lambda}_H)$.

3.2 *Generalized Newton-Raphson estimator*

3.2.1 *Preliminaries.* In order to estimate F^*_{k+1} and H^*_{k+1}, smoothing techniques, as used in (2) and (3), must be utilized in order to borrow strength across subgroups of individuals with differing observed outcome histories. These techniques require the selection of tuning parameters governing the extent of smoothing. As in the above procedure, 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.

If the parameter of interest is a sufficiently smooth functional on the space of possible data-generating distributions, it is sensible to expect a first-order expansion of the form

$$\mu(P) - \mu(P^*) = \int D(P)(o) d(P - P^*)(o) + Rem(P, P^*) \tag{4}$$

to hold, where $D(P)(o)$ is the evaluation at an observation value o of a so-called gradient of μ at P, and $Rem(P, P^*)$ is a second-order remainder term tending to zero as P tends to P^*. In the context of our problem, this is established formally in Lemma 1. Here, the gradient D is an analytic object used to compute, at any given data-generating distribution P, the change in $\mu(P)$ following a slight perturbation of P. Although the gradient is, in general, not uniquely defined, it must have mean zero and finite variance under sampling from P. A discussion on gradients of statistical parameters can be found in Pfanzagl (1982) and in Appendix A.4 of van der Laan and Rose (2011).

Provided (4) holds and for a given estimator $\widehat{P}$ of P^*, algebraic manipulations leads to

$$\begin{aligned}
\mu(\widehat{P}) - \mu(P^*) &= \int D(\widehat{P})(o) d(\widehat{P} - P^*)(o) + Rem(\widehat{P}, P^*) \\
&= \frac{1}{n} \sum_{i=1}^{n} D(P^*)(O_i) + \int \{D(\widehat{P})(o) - D(P^*)(o)\} d(P_n - P^*)(o) \\
&\quad - \frac{1}{n} \sum_{i=1}^{n} D(\widehat{P})(O_i) + Rem(\widehat{P}, P^*) \, ,
\end{aligned}$$

where P_n denotes the empirical distribution based on $O_1, O_2, \ldots, O_n$. If $\widehat{P}$ is a sufficiently well-behaved estimator of P^*, it is often the case that the terms $\int \{D(\widehat{P})(o) - D(P^*)(o)\} d(P_n - P^*)(o)$ and $Rem(\widehat{P}, P^*)$ are asymptotically negligible. However, when $\widehat{P}$ involves smoothing, as in this paper, the term $n^{-1} \sum_{i=1}^{n} D(\widehat{P})(O_i)$ generally tends to zero too slowly to allow $\mu(\widehat{P})$ to be an asymptotically linear estimator of μ^*. Nonetheless, the corrected estimator

$$\widehat{\mu} = \mu(\widehat{P}) + \frac{1}{n} \sum_{i=1}^{n} D(\widehat{P})(O_i)$$

is regular and asymptotically linear with influence function $D(P^*)$, provided that the afore-

mentioned terms are asymptotically negligible. Consequently, $\widehat{\mu}$ converges to μ^* in probability and $n^{1/2}(\widehat{\mu}-\mu^*)$ tends in distribution to a zero-mean Gaussian random variable with variance $\sigma^2 := \int D(P^*)(o)^2 dP^*(o)$. This estimator is, in fact, a direct generalization of the one-step Newton-Raphson procedure used in parametric settings to produce an asymptotically efficient estimator. This correction approach was discussed early on by Ibragimov and Khasminskii (1981), Pfanzagl (1982) and Bickel (1982), among others.

An alternative estimation strategy would consist of employing targeted minimum loss-based estimation (TMLE) to reduce bias due to inadequate tuning (van der Laan and Rubin, 2006). TMLE proceeds by modifying the initial estimator $\widehat{P}$ into an estimator $\widetilde{P}$ that preserves the consistency but also satisfies the equation $n^{-1}\sum_{i=1}^{n} D(\widetilde{P})(O_i) = 0$. As such, the TMLE-based estimator $\widetilde{\mu} := \mu(\widetilde{P})$ of μ^* does not require additional correction and is asymptotically efficient. In preliminary simulation studies (not shown here), we found no substantial difference between the TMLE and our proposed one-step estimator $\widehat{\mu}$. In this case, we favor the latter because of its greater ease of implementation.

3.2.2 *Estimator based on canonical gradient: definition and properties.* In our problem, the one-step estimator can be constructed using any gradient D of the parameter μ defined on the model $\mathscr{M}_0$. Efficiency theory motivates the use of the canonical gradient, often called the efficient influence function, in the construction of the above estimator. The resulting estimator is then not only asymptotically linear but also asymptotically efficient relative to model $\mathscr{M}_0$. The canonical gradient can be obtained by projecting any other gradient onto the tangent space, defined at each $P \in \mathscr{M}_0$ as the closure of the linear span of all score functions of regular one-dimensional parametric models through P. A comprehensive treatment of efficiency theory can be found in Pfanzagl (1982) and Bickel et al. (1993).

In our analysis, we restrict our attention to the class of selection bias functions of the form $\rho_{k+1}(\overline{y}_k, y_{k+1}) = \alpha\rho(y_{k+1})$, where ρ is a specified function of y_{k+1} and α is a sensitivity

analysis parameter. With this choice, $\alpha = 0$ corresponds SI-1. For the parameter chosen, the canonical gradient $D^\dagger(P)$ relative to $\mathscr{M}_0$, suppressing notational dependence on α, is given by

$$D^\dagger(P)(o) := a_0(y_0) + \sum_{k=0}^{K-1} r_{k+1} b_{k+1}(y_{k+1}, y_k) + \sum_{k=0}^{K-1} r_k \{1 - r_{k+1} - H_{k+1}(y_k)\} c_{k+1}(y_k) \ ,$$

where expressions for $a_0(y_0)$, b_{k+1} and c_{k+1} are given the Appendix. In this paper we suggest the use of the following one-step estimator

$$\widehat{\mu} := \mu(\widehat{P}) + \frac{1}{n} \sum_{i=1}^{n} D^\dagger(\widehat{P})(O_i)$$

which stems from linearization (4), as formalized in the following lemma.

LEMMA 1: *For any $P \in \mathscr{M}_0$, the linearization*

$$\mu(P) - \mu(P^*) = \int D^\dagger(P)(o) d(P - P^*)(o) + Rem(P, P^*)$$

holds for a second-order remainder term $Rem(P, P^*)$ defined in Web Appendix B.

In the above lemma, the expression *second-order* refers to the fact that $Rem(P, P^*)$ can be written as a sum of the integral of the product of two error terms each tending to zero as P tends to P^*, that is,

$$Rem(P, P^*) = \sum_{k=0}^{K-1} \int u_k^*(o) \left\{ \Psi_k(P)(o) - \Psi_k(P^*)(o) \right\} \left\{ \Theta_k(P)(o) - \Theta_k(P^*)(o) \right\} dP^*(o) \quad (5)$$

for certain smooth operators $\Psi_0, \ldots, \Psi_{K-1}, \Theta_0, \ldots, \Theta_{K-1}$ and weight functions $u_0^*, \ldots, u_{K-1}^*$ that possibly depend on P^*. The proof of Lemma 1 follows from the derivations in Web Appendices A and B.

The proposed estimator is asymptotically efficient relative to model $\mathscr{M}_0$ under certain regularity conditions, as outlined below.

THEOREM 1: *If (a) $\int \{D^\dagger(\widehat{P})(o) - D^\dagger(P^*)(o)\} d(P_n - P^*)(o) = o_P(n^{-1/2})$ and (b) $Rem(\widehat{P}, P^*) = o_P(n^{-1/2})$, then $\widehat{\mu} = \mu^* + \frac{1}{n} \sum_{i=1}^{n} D^\dagger(P^*)(O_i) + o_P(n^{-1/2})$ and $\widehat{\mu}$ is an asymptotically efficient estimator of μ^* relative to model $\mathscr{M}_0$.*

The proof of this theorem is provided in Web Appendix C. This result not only justifies the use of $\widehat{\mu}$ in practice but also suggests that a Wald-type asymptotic $100 \times (1-\gamma)\%$ confidence interval for μ^* can be constructed as

$$\left(\widehat{\mu} - \frac{z_{\gamma/2}\widehat{\sigma}}{\sqrt{n}},\ \widehat{\mu} + \frac{z_{\gamma/2}\widehat{\sigma}}{\sqrt{n}}\right), \tag{6}$$

where $\widehat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} D^\dagger(\widehat{P})(O_i)^2$ is, under mild conditions, a consistent estimator of the asymptotic variance of $n^{1/2}(\widehat{\mu}-\mu^*)$ and $z_{\gamma/2}$ is the $(1-\gamma/2)$-quantile of the standard normal distribution.

Alternative sufficient conditions can be established to guarantee that conditions (a) and (b) of the theorem above hold. For example, a simple application of Lemma 19.24 of van der Vaart (2000) implies that condition (a) holds provided it can be established that

(i) $D^\dagger(\widehat{P})$ is a consistent estimator of $D^\dagger(P^*)$ in the $L_2(P^*)$-norm in the sense that
$$\int \left\{D^\dagger(\widehat{P})(o) - D^\dagger(P^*)(o)\right\}^2 dP^*(o) \xrightarrow{P} 0, \text{and}$$

(ii) for some P^*-Donsker class $\mathscr{F}$, $D^\dagger(\widehat{P})$ falls in $\mathscr{F}$ with probability tending to one.

Since our estimator $\widehat{P}$ is based on kernel regression, and is therefore consistent, condition (i) holds by a simple application of the continuous mapping theorem. Condition (ii) is standard in the analysis of estimators based on data-adaptive estimation of nuisance parameters – Giné and Nickl (2008) presents conditions under which it is expected to hold. Condition (b) is satisfied based on the following argument. The use of cross-validation allows the optimal rate $n^{-2/5}$ to be achieved for the estimator $\widehat{P}$ since the latter is constructed using univariate kernel smoothers. By a repeated use of the Cauchy-Schwartz inequality on the various summands of $Rem(\widehat{P}, P^*)$ in (5), the continuous mapping theorem allows us to show that, since each term in $Rem(\widehat{P}, P^*)$ is a second-order difference involving smooth transformations of components of $\widehat{P}$ and P, $Rem(\widehat{P}, P^*)$ tends to zero in probability at a rate faster than $n^{-1/2}$ under very mild conditions, including that the probabilities $\widehat{\pi}(Y_{j-1}, Y_j)$ are bounded away from zero with probability tending to one.

3.3 *Practical considerations in confidence interval construction*

As indicated above, an influence function-based asymptotic confidence interval is given by (6). In Section 5, we present the results of a simulation study in which this confidence interval construction results in poor coverage. The poor coverage can be explained in part by the fact that $\widehat{\sigma}^2$ can be severely downward biased in finite samples (Efron and Gong, 1983).

To address this issue, one can consider the jackknife estimator for σ^2,

$$\widehat{\sigma}^2_{JK} := (n-1)\sum_{i=1}^{n}\left(\widehat{\mu}^{(-i)} - \widehat{\mu}^{(\cdot)}\right)^2$$

where $\widehat{\mu}^{(-i)}$ is the estimator of μ^* with the ith individual deleted from the dataset and $\widehat{\mu}^{(\cdot)} := \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}^{(-i)}$. This estimator is known to be conservative (Efron and Stein, 1981). Using the jackknife, confidence intervals take the form of (6) with $\widehat{\sigma}$ replaced by $\widehat{\sigma}_{JK}$. Our simulation study in Section 5 demonstrates that these intervals perform better than interval (6) although some undercoverage is still present.

Another possible approach would be to utilize the Studentized bootstrap, wherein confidence intervals are formed by choosing cutpoints based on the distribution of

$$\left\{\frac{\widehat{\mu}_{(b)} - \widehat{\mu}}{\widehat{se}(\widehat{\mu}_{(b)})} : b = 1, 2, \ldots, B\right\} \tag{7}$$

where $\widehat{\mu}_{(b)}$ is the estimator of μ^* based on the bth bootstrap dataset and $\widehat{se}(\widehat{\mu}_{(b)})$ is an estimator of the standard error of $\widehat{\mu}_{(b)}$. One can consider standard error estimators based on the influence function or jackknife. An equal-tailed $(1-\gamma)$ confidence interval takes the form $\left\{\widehat{\mu} - t_{1-\gamma/2}\widehat{se}(\widehat{\mu}), \widehat{\mu} - t_{\gamma/2}\widehat{se}(\widehat{\mu})\right\}$, where t_q is the qth quantile of (7). A symmetric $(1-\gamma)$ confidence interval takes the form $\left\{\widehat{\mu} - t^*_{1-\gamma}\widehat{se}(\widehat{\mu}), \widehat{\mu} + t^*_{1-\gamma}\widehat{se}(\widehat{\mu})\right\}$, where $t^*_{1-\gamma}$ is selected so that the sampling distribution of (7) assigns probability mass $1-\gamma$ between $-t^*_{1-\gamma}$ and $t^*_{1-\gamma}$.

Since our analysis depends on estimation of a correctly specified semiparametric model, it appears sensible to use this model to bootstrap the observed data. In our data analysis and simulation study, we use the estimated distribution of the observed data to generate boot-

strapped observed datasets. Our simulation study in Section 5 suggests that the symmetric Studentized bootstrap with jackknifed standard errors performs best.

4. SCA-3004 Study

SCA-3004 was a randomized trial designed to evaluate the efficacy and safety of once-monthly, injectable paliperidone palmitate (PP1M), as monotherapy or as an adjunct to pre-study mood stabilizers or antidepressants, relative to placebo (PBO) in delaying the time to relapse in patients with schizoaffective disorder (SCA) (Fu et al., 2014). The study included multiple phases. After initial screening, an open-label phase consisted of a 13-week, flexible-dose, lead-in period and a 12-week, fixed-dose, stabilization period. Stable patients entered a 15-month, double-blind, relapse-prevention phase and were randomized (1:1) to receive either PP1M or placebo injections at baseline (Visit 0) and every 28 days (Visits 1–15). An additional clinic visit (Visit 16) was scheduled 28 days after the last scheduled injection. In the study, 170 and 164 patients were randomized to the PBO and PP1M arms, respectively. One placebo patient was removed because of excessive influence on the analysis.

The main research question was whether or not outcomes in patients with schizoaffective disorder are better maintained if they continued on treatment rather than being withdrawn from treatment and given placebo. Given the explanatory nature of the research question, an ideal study would follow all randomized patients through Visit 16 while maintaining them on their randomized treatment and examine symptomatic and functional outcomes at that time point. Due to ethical considerations, patients who had signs of clinical relapse (determined by symptoms and clinical response to symptoms) were required to be withdrawn from the study. Thus, clinical data were unavailable post-relapse. In addition to this source of missing data, some patients discontinued due to adverse events, withdrew consent or were lost to follow-up. In the trial, 38% and 60% of patients in the PBO and PP1M arms, respectively, were followed through Visit 16 ($p<0.001$).

We focus our analysis on patient function as measured by the Personal and Social Performance (PSP) scale. The PSP, a validated clinician-reported instrument, is scored from 1 to 100, with higher scores indicating better functioning. It has been argued that a clinically meaningful difference in PSP scores is between 7 and 12 points (Patrick et al., 2009).

We seek to estimate, for each treatment group, the mean PSP at Visit 16 in the counterfactual world in which all patients are followed and treated through Visit 16. Since symptoms and function are correlated, the observed PSP data are likely to be a highly biased representation of the counterfactual world of interest. The mean PSP score among completers was 76.53 and 76.96 in the PBO and PP1M arms, respectively; the estimated difference is -0.43 (95% CI: -3.34 to 2.48), indicating a non-significant treatment effect (p=0.77).

In Figure 1, we display the treatment-specific trajectories of mean PSP score, stratified by last visit time. For patients who prematurely terminate the study, it is interesting to notice that there tends to be a worsening of mean PSP scores at the last visit on study.

[Figure 1 about here.]

Before implementing our proposed sensitivity analysis procedure, we implemented the approach of Scharfstein et al. (2014). For each treatment group, we modeled H^*_{k+1} using logistic regression with visit-specific intercepts and a common effect of Y_k. Additionally, we modeled F^*_{k+1} both using beta and truncated normal regression, each with visit-specific intercepts and a common effect of Y_k. Using estimates of the parameters from these models, we simulated 500,000 datasets for each treatment group. We compared the proportion dropping out before visit $k+1$ among those on study at visit k based on the actual and simulated datasets. We also compared the empirical distribution of PSP scores among those on study at visit $k+1$ based on these datasets using the Kolmogorov-Smirnov statistics. The results for the simulations involving the truncated normal regression and beta regression models are shown in the first and second rows of Figure 2, respectively. The figure suggests

that these models do not fit the observed data well. For both the truncated normal and beta regression models, inspection of the actual and simulated distribution of PSP scores at each study visit reveals large discrepancies. For the beta regression model, the contrast between the simulated and actual drop-out probabilities for the PP1M arm is particularly poor.

[Figure 2 about here.]

We contrast the fit of these models to the non-parametric smoothing approach proposed in this paper. For estimation of F^*_{k+1} and H^*_{k+1} based on data from the PBO arm, the optimal choices of λ_F and λ_H are 1.81 and 5.18, respectively. The corresponding optimal choices for the PP1M arm were 1.16 and 8.53. Using the estimated F^*_{k+1} and H^*_{k+1} and optimal choices of λ_F and λ_H, we simulated, as before, 500,000 observed datasets for each treatment group. The results of this simulation in comparison to the actual observed data is shown in the bottom row of Figure 2. In sharp contrast to the parametric modeling approach, the results show excellent agreement between the actual and simulated datasets. For each treatment group, inspection of the actual and simulated distribution of PSP scores at the study visit with the largest Kolmogorov-Smirnov statistics reveals only small discrepancies.

Under SI-1, that is, when $\alpha = 0$, the estimated counterfactual means of interest are 73.31 (95% CI: 69.71 to 76.91) and 74.52 (95% CI: 72.28 to 76.75) for the PBO and PP1M arms, respectively. The estimated treatment difference is -1.20 (95% CI: -5.34 to 2.93). Relative to the complete-case analysis, the SI-1 analysis corrects for bias in a direction that is anticipated: the estimated means under SI-1 are lower and, since there is greater drop-out in the PBO arm, there is a larger correction in that arm. As a consequence, the estimated treatment effect is more favorable to PP1M, although the 95% CI still includes 0. For comparative purposes, the plug-in procedure produces estimates of the means that are slightly lower (73.79 and 74.63) and an estimated treatment difference that is slightly larger (-0.84). The logistic-truncated normal and logistic-beta models for the distribution of the observed data produce markedly

different results under SI-1. For the logistic-truncated model, the estimated means are 70.62 (95% CI: 67.01 to 74.24) and 74.68 (95% CI: 72.89 to 76.48) with an estimated difference of -4.06 (95% CI: -8.13 to 0.01); for the logistic-beta model, the estimated means are 64.42 (95% CI: 55.15 to 73.69) and 70.55 (95% CI: 67.53 to 73.56) with an estimated difference of -6.13 (95% CI: -15.96 to 3.71).

In our sensitivity analysis, we chose ρ to be the cumulative distribution of a $\texttt{Beta}(6, 11)$ random variable scaled to the interval 1 to 100. The shape of the function was chosen so that when comparing patients on the low end ($\leqslant 30$) and high end ($\geqslant 80$) of the PSP scale there is relatively less difference in the risk of drop-out than when comparing patients in the middle of the PSP scale (30-80). When $\alpha > 0$ ($\alpha < 0$), patients with higher PSP scores are more (less) likely to drop out. Since lower PSP scores represent worse function, it is plausible that $\alpha \leqslant 0$. For completeness, we ranged the treatment-specific α values from -20 to 20.

In Figure 3 (a) and (b), we display the estimated treatment-specific mean PSP at Visit 16 as a function of α along with 95% pointwise confidence intervals. Figure 3 (c) displays a contour plot of the estimated differences between mean PSP at Visit 16 for PBO versus PP1M for various treatment-specific combinations of α. The point (0,0) corresponds to the SI-1 assumption in both treatment arms. There are no treatment-specific combinations of α for which the estimated treatment differences are clinically meaningful or statistically significant (at the 0.05 level). Figure 3 (d) displays the estimated treatment-specific difference in mean PSP at Visit 16 between non-completers and completers as a function of α. For each treatment group and α, the estimated mean among non-completers is back-calculated from the estimated overall mean ($\widehat{\mu}$), the observed mean among completers ($\sum_i R_{K,i} Y_{K,i} / \sum_i R_{K,i}$) and the proportion of completers ($\sum_i R_{K,i}/n$). The differences in the negative range of α are in the clinically meaningful range, suggesting that the considered choices of the sensitivity analysis parameters are reasonable.

[Figure 3 about here.]

5. Simulation study

As in our goodness-of-fit evaluation above, we simulated, using the estimated F_k^* and H_k^* and optimal choices of λ_F and λ_H, 1,000 datasets for each treatment group. For purposes of the simulation study, we treat the best fit to the observed data as the true data generating mechanism. We evaluate the performance of our procedures for various α values ranging from -10 to 10. The target for each α is the mean computed using formula (1).

The results of our simulation study are displayed in Tables 1 and 2. In Table 1, we report for each treatment group and each α the bias and mean-squared error (MSE) for the plug-in estimator $\mu(\widehat{P})$ and the one-step estimator $\widehat{\mu}$. The results show that the one-step estimator has less bias and lower MSE than the plug-in estimator, although the differences are not dramatic. In Table 2, we report, for each treatment group and each α, 95% confidence interval coverage for six confidence interval procedures: (1) normality-based confidence interval with influence function-based standard error estimator (Normal-IF); (2) normality-based confidence interval with jackknife-based standard error estimator (Normal-JK); (3) equal-tailed, Studentized-t bootstrap confidence interval with influence function-based standard error estimator (Bootstrap-IF-ET); (4) equal-tailed, Studentized-t bootstrap confidence interval with jackknife-based standard error estimator (Bootstrap-JK-ET); (5) symmetric, Studentized-t bootstrap confidence interval with influence function-based standard error estimator (Bootstrap-IF-S); (6) symmetric, Studentized-t bootstrap confidence interval with jackknife-based standard error estimator (Bootstrap-JK-S). Bootstrapping was based on 1,000 datasets.

[Table 1 about here.]

[Table 2 about here.]

We found that the normality-based confidence interval with influence function-based standard error estimator underperformed for both treatment groups and all choices of the sensitivity analysis parameters. In general, the confidence interval procedures that used jackknife standard errors performed better than their counterparts that used the influence function-based standard error estimator. The symmetric, Studentized-t bootstrap confidence interval with jackknife-based standard error estimator (Bootstrap-JK-S) exhibited the most consistent performance across treatment groups and sensitivity analysis parameters.

Our simulation studies reveal some evidence of possible residual bias of the one-step estimator in the context considered. The latter is based upon the use of kernel smoothing in order to estimate the various conditional distribution functions required in the evaluation of μ. It may be possible to achieve better small-sample behavior by employing alternative conditional distribution function estimators with better theoretical properties, e.g., Hall et al. (1999). An ensemble learning approach, e.g., van der Laan et al. (2007), may also yield improved function estimators and decrease the residual bias of the resulting one-step estimator. However, the benefits from improved function estimation may possibly be limited by the relatively small sample size investigated in this simulation study. The use of correction procedures based on higher-order asymptotic representations, as described in Robins et al. (2008), van der Vaart et al. (2014), Carone et al. (2014) and Díaz et al. (2016), may lead to improved performance in smaller samples.

6. Discussion

In this paper, we have developed a semi-parametric method for conducting a global sensitivity analysis of repeated measures studies with monotone missing data. We have developed an open-source software package, called `SAMON`, that implements the methods discussed in this paper.

Our approach does not, as of yet, accommodate auxiliary covariates V_k scheduled to be

measured at assessment k. Incorporating $\overline{V}_k$ into the conditioning arguments of Assumptions 1 and 2 can serve to increase the plausibility of these assumptions. In particular, $\overline{V}_k$ can be allowed to influence the decision, for patients on study at visit k, to drop out between visits k and $k+1$, and the unmeasured common causes of $Y_0, Y_1, \ldots, Y_K$ can be allowed to indirectly impact the decision to drop out through their relationship with $\overline{V}_k$. In the context of SCA-3004, it would be useful to incorporate the PANSS (Positive and Negative Symptom Scale) and CGI (Clinical Global Impressions) scores as auxiliary covariates as they are related to planned patient withdrawal as well as correlated with PSP. In future work, we plan to extend the methods developed here to accommodate auxiliary covariates. An extension that handles multiple reasons for drop-out is also worthwhile.

In this paper, we imposed a first-order Markovian assumption in modeling the distribution of the observed data. The plausibility of this assumption was considered in the data analysis as we have evaluated the goodness-of-fit of our model, as illustrated in the bottom row of Figure 2. The Markovian assumption can be relaxed by incorporating the past history using (1) a specified function of the past history, (2) semiparametric single index models (Hall and Yao, 2005) or (3) recently developed methods in data adaptive non-parametric function estimation (van der Laan, 2015).

For given α, our estimator of μ^* is essentially an α-specific weighted average of the observed outcomes at visit K. As a result, it does not allow extrapolation outside the support of these outcomes. We found that one patient in the PBO arm who completed the study with the lowest observed PSP score at the final visit had a very large influence on the analysis. Under SI-1 and other values of α, this patient affected the estimated mean in the PBO group by more than 3 points. In contrast to our approach, a mixed modeling approach, which posits a multivariate normal model for the joint distribution of the full data, does allow extrapolation. Inference under this approach is valid under MAR and correct

specification of the multivariate normality assumption. We found that this approach provides much more precise inference, yielding a statistically significant treatment effect in favor of PP1M (treatment effect = -4.7, 95% CI: -7.7 to -1.8). Further, this approach was insensitive to the PBO patient that we removed from our analysis. The disadvantages of the mixed model approach are its reliance on normality and the difficulty of incorporating it into global sensitivity analysis.

In SCA-3004 there is a difference, albeit not statistically significant, in baseline PSP score between treatment groups. The PBO arm has a lower baseline mean PSP score than the PP1M arm (71.2 vs. 72.9). Our method can easily address this imbalance by subtracting out this difference from our effect estimates or by formally modeling change from baseline. In either case, the treatment effect estimates would be less favorable to PP1M. It is notable that a mixed model analysis that models change from baseline does yield a statistically significant effect in favor of PP1M. It may also be of interest to adjust the treatment effect estimates for other baseline covariates, either through regression or direct standardization. We will address this issue in future work. We also plan to develop methods for handling intermittent missing outcome data.

7. Supplementary Materials

Web Appendices referenced in Section 3.2.2 are available with this paper at the Biometrics website on Wiley Online Library. The software package `SAMON` can be found at `www.missingdatamatters.org`.

Acknowledgments and Conflicts

This research was sponsored by contracts from the U.S. Food and Drug Administration and the Patient Centered Outcomes Research Institute as well as NIH grant CA183854. The first and second authors (DS and AM) have received compensation from Janssen Research and

Development, LLC for the provision of consulting services; they received no compensation for preparation of this manuscript or the methods contained herein.

References

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics* pages 647–671.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society: Series B* **65,** 275–297.

Carone, M., Díaz, I., and van der Laan, M. J. (2014). Higher-order targeted minimum loss-based estimation. Technical report, University of California Berkeley, Department of Biostatistics.

CHMP (2009). *Guideline on Missing Data in Confirmatory Clinical Trials*. EMEA, London.

Díaz, I., Carone, M., and van der Laan, M. J. (2016). Second-order inference for the mean of a variable missing at random. *The International Journal of Biostatistics* **12,** 333–349.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37,** 36–48.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics* pages 586–596.

Fu, D.-J., Turkoz, I., Simonson, R. B., Walling, D. P., Schooler, N. R., Lindenmayer, J.-P., Canuso, C. M., and Alphs, L. (2014). Paliperidone palmitate once-monthly reduces risk of relapse of psychotic, depressive, and manic symptoms and maintains functioning in a double-blind, randomized study of schizoaffective disorder. *The Journal of Clinical Psychiatry* **76,** 253–262.

Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields* **141,** 333–387.

Hall, P., Wolff, R. C., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94,** 154–163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Annals of Statistics* pages 1404–1421.

Ibragimov, I. A. and Khasminskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer.

ICH (1998). *Statistical Principles for Clinical Trials (E9)*. Geneva.

Linero, A. R. and Daniels, M. J. (2015). 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,** 45–55.

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. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press.

Little, R. J. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Patrick, D. L., Burns, T., Morosini, P., Rothman, M., Gagnon, D. D., Wild, D., and Adriaenssen, I. (2009). Reliability, validity and ability to detect change of the clinician-rated personal and social performance scale in patients with acute symptoms of schizophrenia. *Current Medical Research and Opinion* **25,** 325–338.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. Springer.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). Higher-order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics:*

Essays in Honor of David A. Freedman, pages 335–421. Institute of Mathematical Statistics.

Robins, J. M., Ritov, Y., et al. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16,** 285–319.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. CRC Press.

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

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data. 2006*. Springer Verlag, New York.

van der Laan, M. (2015). A generally efficient targeted minimum loss based estimator. Technical report, University of California Berkeley, Department of Biostatistics.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2,** Article 11.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6,**.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.

van der Vaart, A. et al. (2014). Higher-order tangent spaces and influence functions.

Statistical Science **29,** 679–686.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W., Dudoit, S., and van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions* **24,** 351–371.

Appendix: Canonical Gradient

The derivation of the canonical gradient is provided in Web Appendix A. Here, we present its explicit form. Let $\pi_{k+1}(y_k, y_{k+1}) = [1 + \exp\{\ell_{k+1}(y_k) + \alpha\rho(y_{k+1})\}]^{-1}$, where

$$\ell_{k+1}(y_k) := \text{logit}\left\{H_{k+1}(y_k)\right\} - \log\left\{\int \exp\{\rho_{k+1}(\overline{y}_k, u)\} dF_{k+1}(u \mid y_k)\right\}.$$

Let $\pi(\overline{y}_K) = \prod_{k=0}^{K-1} \pi_k(y_k, y_{k+1})$,

$$w_{k+1}(y_k) = E\left(\exp\{\alpha\rho(Y_{k+1})\} \mid R_{k+1} = 1, Y_k = y_k\right),$$

and $g_{k+1}(y_{k+1}, y_k) = \{1 - H_{k+1}(y_k)\} w_{k+1}(y_k) + \exp\{\alpha\rho(y_{k+1})\} H_{k+1}(y_k)$.

The canonical gradient is expressed as

$$D^\dagger(P)(o) := a_0(y_0) + \sum_{k=0}^{K-1} r_{k+1} b_{k+1}(y_{k+1}, y_k) + \sum_{k=0}^{K-1} r_k \{1 - r_{k+1} - H_{k+1}(y_k)\} c_{k+1}(y_k)$$

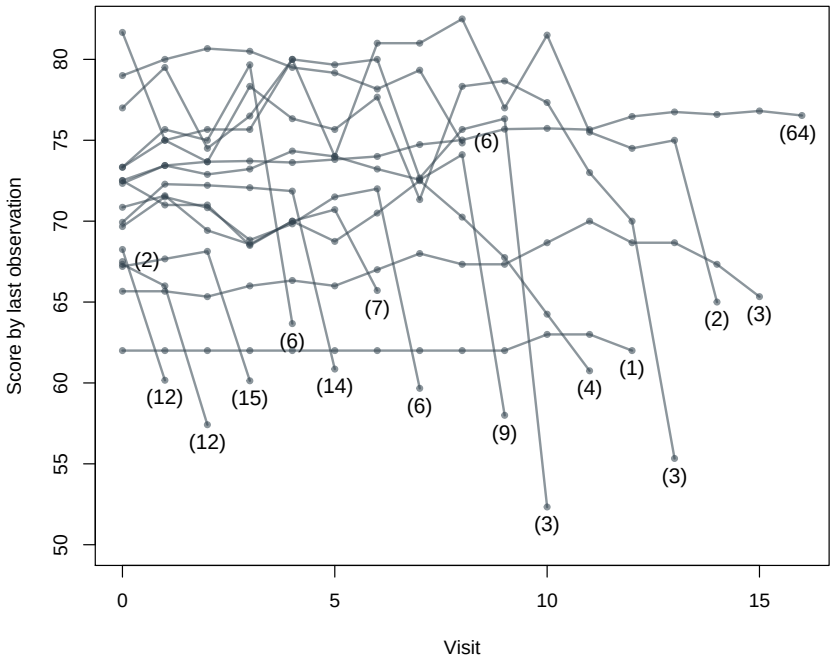
where

$$a_0(y_0) \;=\; E\left(\frac{R_K Y_K}{\pi(\overline{Y}_K)} \,\middle|\, Y_0 = y_0\right) - \mu(P)$$

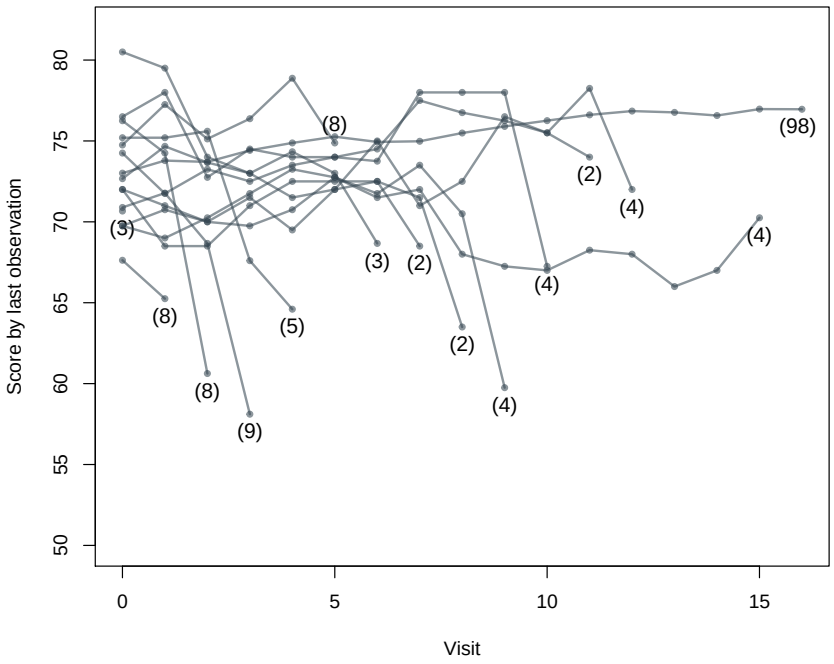
$$\begin{aligned} & b_{k+1}(y_{k+1}, y_k) \\ &= E\left(\frac{R_K Y_K}{\pi(\overline{Y}_K)} \,\middle|\, R_{k+1} = 1, Y_{k+1} = y_{y+1}, Y_k = y_k\right) - E\left(\frac{R_K Y_K}{\pi(\overline{Y}_K)} \,\middle|\, R_{k+1} = 1, Y_k = y_k\right) \\ &\quad + E\left(\frac{R_K Y_K}{\pi(\overline{Y}_K)} \left(\frac{\exp\{\alpha\rho(Y_{k+1})\}}{g_{k+1}(Y_{k+1}, Y_k)}\right) \,\middle|\, R_{k+1} = 1, Y_k = y_k\right) H_{k+1}(y_k) \left(1 - \frac{\exp\{\alpha\rho(y_{k+1})\}}{w_{k+1}(y_k)}\right) \end{aligned}$$

$$\begin{aligned} & c_{k+1}(y_k) \\ &= \;\; E\left(\frac{R_K Y_K}{\pi(\overline{Y}_K)} \left(\frac{\exp\{\alpha\rho(Y_{k+1})\}}{g_{k+1}(Y_{k+1}, Y_k)}\right) \,\middle|\, R_k = 1, Y_k = y_k\right) \\ &\quad - E\left(\frac{R_K Y_K}{\pi(\overline{Y}_K)} \left(\frac{1}{g_{k+1}(Y_{k+1}, Y_k)}\right) \,\middle|\, R_k = 1, Y_k = y_k\right) w_{k+1}(y_k) \end{aligned}$$

Figure 1: Treatment-specific trajectories of mean PSP scores, stratified by last visit time.



(a) Placebo



(b) PP1M

Figure 2: Left column: Comparison of the proportion dropping out before visit $k+1$ among those on study at visit k based on the actual and simulated datasets. Right column: Comparison, using the Kolmogorov-Smirnov statistics, of the empirical distribution of PSP scores among those on study at visit $k+1$ based on the actual and simulated datasets. First row: Logistic regression for conditional probabilities of drop-out and truncated normal regressions for outcomes; Second row: Logistic regression for conditional probabilities of drop-out and beta regressions for outcomes; Third row: Non-parametric smoothing for conditional probabilities of drop-out and for outcomes.

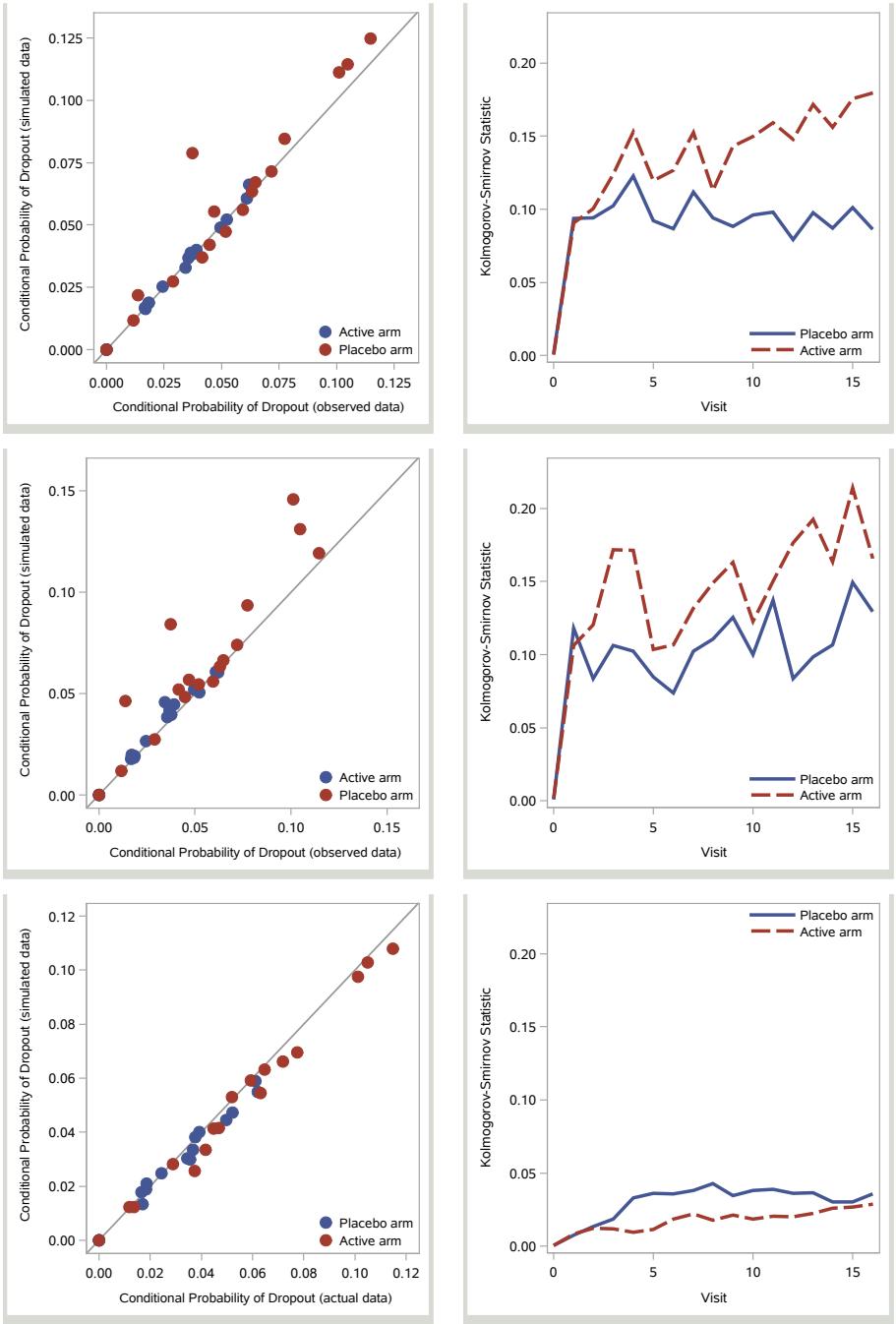


Figure 3: (a) and (b): Treatment-specific mean PSP at Visit 16 as a function of α, along with 95% pointwise confidence intervals; (c): Contour plot of the estimated differences between mean PSP at Visit 16 for PBO vs. PP1M for various treatment-specific combinations of α; (d): Treatment-specific differences between the mean PSP for non-completers and completers, as a function of α.

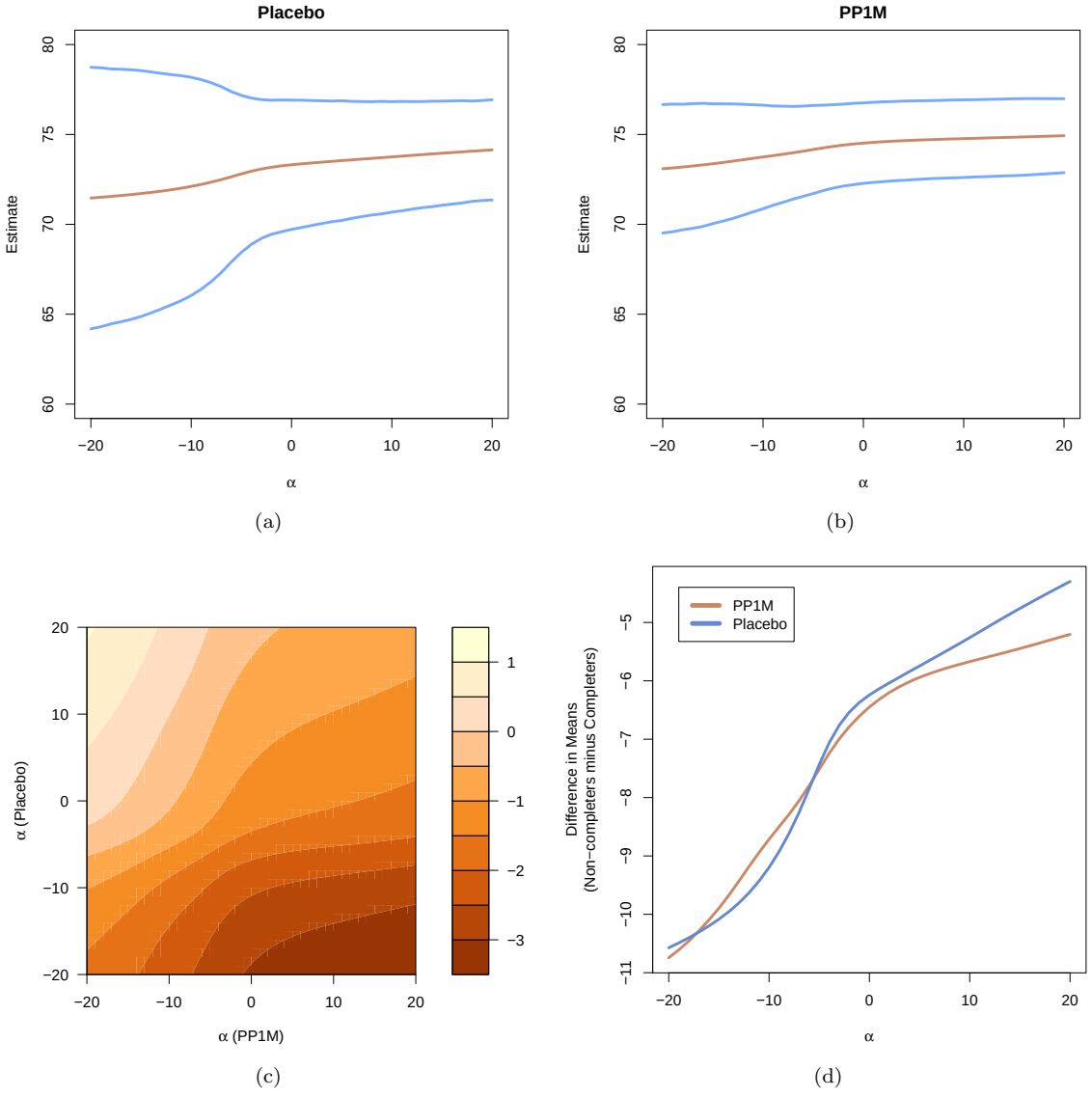


Table 1: Treatment-specific simulation results: Bias and mean-squared error (MSE) for the plug-in ($\mu(\widehat{P})$) and one-step ($\widehat{\mu}$) estimators, for various choices of α.

		PBO			PP1M		
α	Estimator	μ^*	Bias	MSE	μ^*	Bias	MSE
-10	$\mu(\widehat{P})$	72.89	0.76	1.75	73.76	0.41	1.36
	$\widehat{\mu}$		0.50	1.58		0.31	1.26
-5	$\mu(\widehat{P})$	73.38	0.52	1.42	74.25	0.26	1.14
	$\widehat{\mu}$		0.31	1.32		0.16	1.05
-1	$\mu(\widehat{P})$	73.74	0.38	1.23	74.59	0.17	1.02
	$\widehat{\mu}$		0.19	1.18		0.06	0.95
0	$\mu(\widehat{P})$	73.80	0.36	1.21	74.63	0.16	1.01
	$\widehat{\mu}$		0.18	1.17		0.08	0.95
1	$\mu(\widehat{P})$	73.84	0.35	1.19	74.67	0.18	1.01
	$\widehat{\mu}$		0.17	1.15		0.05	0.94
5	$\mu(\widehat{P})$	74.00	0.30	1.13	74.67	0.16	1.00
	$\widehat{\mu}$		0.13	1.11		0.04	0.93
10	$\mu(\widehat{P})$	74.15	0.24	1.08	74.84	0.15	0.97
	$\widehat{\mu}$		0.10	1.08		0.06	0.91

Table 2: Treatment-specific simulation results: Coverage for (1) normality-based confidence interval with influence function-based standard error estimator (Normal-IF); (2) normality-based confidence interval with jackknife-based standard error estimator (Normal-JK); (3) equal-tailed, Studentized-t bootstrap confidence interval with influence function-based standard error estimator (Bootstrap-IF-ET); (4) equal-tailed, Studentized-t bootstrap confidence interval with jackknife-based standard error estimator (Bootstrap-JK-ET); (5) symmetric, Studentized-t bootstrap confidence interval with influence function-based standard error estimator (Bootstrap-IF-S); (6) symmetric, Studentized-t bootstrap confidence interval with jackknife-based standard error estimator (Bootstrap-JK-S), for various choices of α.

		PBO	PP1M
α	Procedure	Coverage (%)	Coverage (%)
-10	Normal-IF	86.1	88.6
	Normal-JK	92.1	92.6
	Bootstrap-IF-ET	90.2	91.9
	Bootstap-JK-ET	92.4	93.7
	Bootstap-IF-S	92.3	92.7
	Bootstap-JK-S	93.9	94.3
-5	Normal-IF	89.0	91.7
	Normal-JK	94.1	94.2
	Bootstrap-IF-ET	91.7	92.6
	Bootstap-JK-ET	93.6	94.9
	Bootstap-IF-S	94.1	94.2
	Bootstap-JK-S	95.1	95.1
-1	Normal-IF	90.8	93.4
	Normal-JK	94.9	94.8
	Bootstrap-IF-ET	91.0	94.0
	Bootstap-JK-ET	92.8	94.9
	Bootstap-IF-S	94.4	94.7
	Bootstap-JK-S	95.0	95.3
0	Normal-IF	90.7	93.5
	Normal-JK	95.0	94.9
	Bootstrap-IF-ET	92.8	93.9
	Bootstap-JK-ET	94.3	95.0
	Bootstap-IF-S	95.3	94.7
	Bootstap-JK-S	96.0	95.1
1	Normal-IF	90.9	93.5
	Normal-JK	94.9	94.8
	Bootstrap-IF-ET	92.8	93.5
	Bootstap-JK-ET	94.2	95.0
	Bootstap-IF-S	95.3	94.6
	Bootstap-JK-S	96.0	95.2
5	Normal-IF	91.5	93.7
	Normal-JK	94.6	95.1
	Bootstrap-IF-ET	92.6	93.8
	Bootstap-JK-ET	93.8	94.7
	Bootstap-IF-S	94.9	95.1
	Bootstap-JK-S	96.0	95.5
10	Normal-IF	92.1	93.4
	Normal-JK	94.8	95.0
	Bootstrap-IF-ET	92.9	93.8
	Bootstap-JK-ET	93.9	94.8
	Bootstap-IF-S	94.7	95.0
	Bootstap-JK-S	95.6	95.4