Biometrics 000, 000–000 DOI: 000
000 0000

Global Sensitivity Analysis of Randomized Trials with Non-Monotone Missing Binary Outcomes: Application to Studies of Substance Use Disorders

Daniel O. Scharfstein[1,*], Jon Steingrimsson[2], Aidan McDermott[1], Chenguang Wang[3], Souvik Ray[4], Aimee Campbell[5], Edward Nunes[5] and Abigail Matthews[6]

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

[2]Department of Biostatistics, Brown University, Providence, RI 02903

[3]Division of Biostatistics and Bioinformatics, The Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205

[4]Department of Statistics, Stanford University, Stanford, CA 94305

[5]Department of Psychiatry, Columbia University Medical Center and New York State Psychiatric Institute, New York, NY 10032

[6]The EMMES Corporation, Rockville, MD 20850

email: dscharf@jhu.edu

Summary: In this paper, we present a method for conducting global sensitivity analysis of randomized trials in which binary outcomes are scheduled to be collected on participants at fixed in time after randomization and these outcomes may be missing in a non-monotone fashion. We introduce a class of missing data assumptions, indexed by sensitivity parameters, that are anchored around the missing not at random assumption introduced by Robins (*Statistics in Medicine*, 1997). For each assumption in the class, we establish that the joint distribution of the outcomes are identifiable from the distribution of the observed data. Our estimation procedure uses the plug-in principle, where the distribution of the observed data is estimated using random forests. We establish $\sqrt{n}$ asymptotic properties for our estimation procedure. We illustrate our methodology in the context of a randomized trial designed to evaluate a new approach to reducing substance use, assessed by testing urine samples twice weekly, among patients entering outpatient addiction treatment. We evaluate the finite sample properties of our method in a realistic simulation study. Our methods have been implemented in an `R` package entitled `slabm`.

Key words: Missing Not at Random; Random Forests

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

1. Introduction

Missing outcome data threaten the validity of randomized clinical trials because inference about treatment effects must then rely on untestable assumptions. As a result, the National Research Council (NRC) in its report entitled "The Prevention Treatment of Missing Data in Clinical Trials" recommended that evaluating the sensitivity of trial results to assumptions about the missing data mechanism should be a mandatory component of reporting (Little et al., 2010). There does not appear to be concensus, however, about what constitutes an adequate sensitivity analysis. Chapter 5 of the NRC Report presents an approach whereby one posits a broad class of untestable missing data assumptions that are: (1) indexed by sensitivity analysis parameters, (2) anchored around a plausible benchmark assumption (sensitivity parameters equal to a reference value), and (3) sensitivity analysis parameters further from the reference value represent larger deviations from the benchmark assumption (Little et al., 2010). The goal of this "global" sensitivity analysis approach is to determine how much deviation from a benchmark assumption is required in order for 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.

Randomized trials are frequently designed so participants are scheduled to have assessments at fixed points in time after randomization. For participants who miss one or more assessments, their missing data pattern can be classified as either "monotone" or "non-monotone": if all assessments are missing after the first missing visit then the pattern is monotone, otherwise it is non-monotone (i.e., missing visits interspersed about completed visits). Positing plausible assumptions and specifying flexible models for studies with non-monotone missing data is challenging because of the potentially large number of missingness patterns (as many as $2^K - 1$ patterns, where K is the number of post-baseline assessments).

Ibrahim and Molenberghs (2009) indicate that "[s]uch data present a considerable modeling challenge for the statistician". The NRC report highlighted the need for development and application of "novel, appropriate methods of model specification and sensitivity analyses to handle non-monotone missing data patterns" (Little et al., 2010).

1.1 *Identification Assumptions for Non-Monotone Missing Data*

One of the most common assumptions used to identify treatment effects in longitudinal studies is the missing at random (MAR) assumption. This untestable assumption states that, for each possible missingness pattern, the probability of observing the pattern does not depend on the unobserved outcomes conditional on the outcomes that are observed. A convenient feature of this assumption is that if one adopts a likelihood or Bayesian inferential perspective, one just needs to specify a fully parametric model for the joint distribution of the outcomes; the conditional probability of the missingness pattern given the outcomes factors out of the likelihood and can be ignored. This is why many studies are analyzed using mixed models.

While MAR has been considered a reasonable benchmark assumption for studies that have monotone missing data patterns, Robins (1997) and Little and Rubin (2014) have argued that MAR is implausible for studies that have non-monotone missing data patterns. Minini and Chavance (2004) and Fitzmaurice et al. (2018) developed likelihood-based global sensitivity analysis procedures for non-monotone missing binary outcomes anchored at MAR. Alternative assumptions have been proposed:

A1: Little (1993) introduced the complete case missing value (CCMV) assumption. CCMV posits that, for each possible pattern with missing observations, the conditional distribution of the missing outcomes given the observed outcomes is equal to corresponding distribution for the pattern with no missing observations. Tchetgen-Tchetgen et al. (2017) developed a global sensitivity analysis procedure anchored at CCMV.

A2: Vansteelandt et al. (2007) assumed that, for individuals who have the same observed data prior to a scheduled visit, the distribution of the outcome for those missing the visit is the same as the distribution of the outcome for those who attend the visit. They developed a global sensitivity analysis procedure anchored at this assumption. Linero and Daniels (2018) built a Bayesian synthesis procedure.

A3: Zhou et al. (2010) assumed that, for individuals who share the same outcomes (observed or not) and same missingness pattern prior to a scheduled visit, the distribution of the outcome for those missing the visit is the same as the distribution of the outcome for those who attend the visit. No global sensitivity analysis procedure was developed.

A5: Sadinle and Reiter (2017) and Shpitser (2016) assumed that, for individuals who have the same outcomes (observed or not) and same missingness pattern prior to and after a scheduled visit, the distribution of the outcome for those missing the visit is the same as the distribution of the outcome for those who attend the visit. Sadinle and Reiter (2017) developed a Bayesian global sensitivity analysis procedure anchored at this assumption.

A6: Robins (1997) and Sadinle and Reiter (2018) assumed that, for individuals who share the same outcomes (observed or not) prior to a scheduled visit and the same observed data after the visit, the distribution of the outcome for those missing the visit is the same as the distribution of the outcome for those attending the visit. No global sensitivity analysis procedure was developed.

As noted by Little (1993), A1 is not a plausible benchmark assumption. For example, it unrealistically assumes that the joint distribution of outcomes for individuals who miss all their assessments is equal to the joint distribution of outcomes for individuals who complete all their assessments. A2, A3 and A4 are not satisfactory as they do not allow the distribution of the missing outcome at a visit to depend on all the observable factors (e.g., outcomes that are observed after the missing visit). The problem with A5 is that it cannot be represented

in terms of a directed acyclic graph (DAG), which is fundamental for describing a data generating process. Rather A5 can be represented as chain graph, which can be very difficult to interpret (Lauritzen and Richardson, 2002). A7 can be represented as a DAG and does allow the distribution of the missing outcome at a visit to depend on outcomes that are observed after the missing visit. The statistical innovation of this paper is to develop, for binary outcomes, a flexible global sensitivity analysis procedure anchored at A6.

1.2 *Substance Use Disorder Studies*

Our methods are motivated by randomized trials that evaluate treatments for substance use disorders. Trials of individuals with such disorders are well known to suffer from high rates of missing outcome data (Yang and Shoptaw, 2005; McPherson et al., 2012). In substance use disorder trials, there can be high rates of non-monotone missing data.

Consider CTN-0044, a randomized trial designed to evaluate a new approach to reducing substance use among patients entering outpatient addiction treatment (Campbell et al., 2014). In this study, patients were randomized to 12 weeks of either treatment-as-usual (TAU, $n = 252$) or treatment-as-usual plus a computerized therapeutic education system and contingent incentives (TAU+, $n = 255$). Urine samples were scheduled to be collected twice weekly. Focusing on the first 6 weeks, 67.9% and 60.0% of individuals randomized to the TAU and TAU+ arms had non-monotone missing data patterns, respectively.

1.3 *Outline of Paper*

The paper is organized as follows. In Section 2, we introduce the data structure and notation. In Section 3, we introduce our class of missing data assumptions, indexed by sensitivity analysis. We present a theorem that shows that the joint distribution of the outcomes in a world without missing data is identified for each member of the class of missing data assumptions. In Section 4, we consider how to draw inference in finite samples. We propose to model the distribution of the observed data using random forests and estimate functionals

of the distribution of interest using the plug-in principle. Section 5 presents a re-analysis of CTN-0044. The last section is devoted to a discussion.

2. Data Structure and Notation

We consider a trial in which a binary outcome (e.g., substance use) is scheduled to be measured at K post-baseline clinic visits. Let $Y_k^{(1)}$ denote the binary outcome at visit k ($k = 1, \ldots, K$). Let R_k be the binary indicator that $Y_k^{(1)}$ is observed. Let $Y_k = Y_k^{(1)}$ if $R_k = 1$ and $Y_k =?$ if $R_k = 0$. Let O_k be the observed data at visit k; it can be represented by (R_k, Y_k) or by the two-dimensional vector $O_k = (I(R_k = 1, Y_k = 1), I(R_k = 1, Y_k = 0))$. For a time varying quantity Z_k, let $\overleftarrow{Z}_k = (Z_1, \ldots, Z_k)$ and $\overrightarrow{Z}_k = (Z_{k+1}, \ldots, Z_K)$. Assume that we observe n independent and identically distributed copies for $\overleftarrow{O}_K$. The goal is to use the observed data to draw inference about a feature of the distribution of $\overleftarrow{Y}_K^{(1)}$. This feature is the target parameter of interest.

Below, we define $P_1(o_1) = P(O_1 = o_1)$ and, for $k = 2, \ldots, K$, $P_k(\overleftarrow{o}_k) = P_k(\overleftarrow{O}_k = \overleftarrow{o}_k)$ and $P_k(o_k|\overleftarrow{o}_{k-1}) = P(O_k = o_k|\overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1})$. We use $\widehat{P}_1(o_1)$, $\widehat{P}_k(\overleftarrow{o}_k)$ and $\widehat{P}_k(o_k|\overleftarrow{o}_{k-1})$ to denote the corresponding estimators.

3. Assumptions and Identification

We build a novel class of missing data assumptions using the exponential tilting device (Cox and Barndorff-Nielsen, 1994). Imagine a stratum of individuals who share the same substance use history prior to visit k and same observed data after visit k. Now, imagine splitting the stratum into two sets: those who provide outcome data at visit k (stratum A) and those who do not (stratum B). We assume

$$P(Y_k^{(1)} = 1|\underbrace{R_k = 0, \overleftarrow{Y}_{k-1}^{(1)}, \overrightarrow{O}_k}_{\text{Stratum B}}) \propto P(Y_k^{(1)} = 1|\underbrace{R_k = 1, \overleftarrow{Y}_{k-1}^{(1)}, \overrightarrow{O}_k}_{\text{Stratum A}}) \exp(\alpha_k) \qquad (1)$$

where α_k is the sensitivity analysis parameter. When $\alpha_k > 0$ (< 0), it is assumed that stratum B individuals are more (less) likely to have $Y_k^{(1)} = 1$ than stratum A individuals. As $\alpha_k \to \infty$ ($-\infty$), it is assumed that all individuals in stratum B have $Y_k^{(1)} = 1$ ($Y_k^{(1)} = 0$). Notice that when $\alpha_k = 0$ for all k, the benchmark assumption A6 is obtained.

Figure 1 presents a directed acyclic graph (DAG) representation of our assumptions for the case when $K = 4$. In this DAG, there are arrows into Y_k from $Y_k^{(1)}$ and R_k since Y_k is a deterministic function of these latter variables. The red arrow from $Y_k^{(1)}$ into R_k represents the dependence implied by the sensitivity analysis parameter α_k. The red arrows will be absent when $\alpha_k = 0$ for all k.

[Figure 1 about here.]

THEOREM 1: *The distribution of* $\overleftarrow{Y}_K^{(1)}$ *is identified under Assumption (1) for specified* $\overleftarrow{\alpha}_K$.

Proof. By mathematical induction, we can show that

$$P(\overleftarrow{Y}_k^{(1)} = \overleftarrow{y}_k^{(1)}, \overrightarrow{O}_k = \overrightarrow{o}_k) \tag{2}$$

is identifiable for all k, $\overleftarrow{y}_k^{(1)}$ and $\overrightarrow{o}_k$. For $k = 0$, identification is trivial since (2) is equal to $P(O_1 = o_1, \ldots, O_K = o_K)$. Suppose that (2) is identified for $k = s - 1$ ($s \geqslant 1$) - induction hypothesis. We need to prove that it is identfied for $k = s$. By the law of total probability,

$$P(\overleftarrow{Y}_s^{(1)} = \overleftarrow{y}_s^{(1)}, \overrightarrow{O}_s = \overrightarrow{o}_s) = \sum_{j=0}^{1} P(\overleftarrow{Y}_{s-1}^{(1)} = \overleftarrow{y}_{s-1}^{(1)}, Y_s = y_s, R_s = j, \overrightarrow{O}_s = \overrightarrow{o}_s)$$

Now, $P(\overleftarrow{Y}_{s-1}^{(1)} = \overleftarrow{y}_{s-1}^{(1)}, Y_s^{(1)} = y_s^{(1)}, R_s = 1, \overrightarrow{O}_s = \overrightarrow{o}_s)$ is equal to $P(\overleftarrow{Y}_{s-1}^{(1)} = \overleftarrow{y}_{s-1}^{(1)}, \overrightarrow{O}_{s-1} = \overrightarrow{o}_{s-1})$ with $(O_s = o_s) = (R_s = 1, Y_s^{(1)} = y_s^{(1)})$, which is identified by the induction hypothesis. To complete the proof, it is sufficient to establish identification of $P(\overleftarrow{Y}_{s-1}^{(1)} = \overleftarrow{y}_{s-1}^{(1)}, Y_s^{(1)} = y_s^{(1)}, R_s = 0, \overrightarrow{O}_s = \overrightarrow{o}_s)$, which can be written as

$$\underbrace{P(Y_s^{(1)} = y_s^{(1)} | R_s = 0, \overleftarrow{Y}_{s-1}^{(1)} = \overleftarrow{y}_{s-1}^{(1)}, \overrightarrow{O}_s = \overrightarrow{o}_s)}_{\text{Term 1}} \underbrace{P(\overleftarrow{Y}_{s-1}^{(1)} = \overleftarrow{y}_{s-1}^{(1)}, R_s = 0, \overrightarrow{O}_s = \overrightarrow{o}_s)}_{\text{Term 2}}$$

Term 2 is equal to $P(\overleftarrow{Y}^{(1)}_{s-1} = \overleftarrow{y}^{(1)}_{s-1}, \overrightarrow{O}_{s-1} = \overrightarrow{o}_{s-1})$ with $(O_s = o_s) = (R_s = 0)$, which is identified by the induction hypothesis. Term 1 is identified since it can be determined by right hand side of (1), which is itself identified by the induction hypothesis. By setting $k = K$ in (2), we obtain identification of $P(\overleftarrow{Y}^{(1)}_{K} = \overleftarrow{y}^{(1)}_{K})$.

4. Inference

To estimate the $P(\overleftarrow{Y}^{(1)}_{K} = \overleftarrow{y}^{(1)}_{K})$ for specified $\overleftarrow{\alpha}_K$, we estimate the distribution of $\overleftarrow{O}_K$ and use it as a *plug-in* into the identification procedure described in the proof of Theorem 1. To proceed, note that

$$P_K(\overleftarrow{o}_K) = P_1(o_1) \prod_{k=2}^{K} P_k(o_k|\overleftarrow{o}_{k-1}) \tag{3}$$

To estimate the joint distribution of $\overleftarrow{O}_K$, we form the product of separate estimates of each distribution on the right hand side of (3). Specifically, we estimate the distribution O_1 by its empirical distribution and then estimate O_k given $\overleftarrow{O}_{k-1}$, for $k = 2, \ldots, K$ using random forests. Briefly, the random forest algorithm is built on top of the classification and regression tree (CART) algorithm, which creates a risk prediction model by recursively partitioning the covariate space $\overleftarrow{O}_{k-1}$ using binary splits (Breiman et al., 1984). With ternary outcomes (O_k), the decision to split is made by minimizing a measure of impurity (e.g., Gini impurity). For fully grown CART trees splitting is continued until each terminal node has at most D observations, for a pre-determined integer D.

With the aim of improving prediction accuracy, Breiman (1996) proposed an ensemble algorithm, referred to as bagging, that averages fully grown CART trees built using different bootstrap samples. To de-correlate the individual trees in the ensemble, the random forest algorithm (Breiman, 2001) modifies bagging by only considering a subset of the covariates at each splitting decision.

Let $\widehat{P}_K(\overleftarrow{o}_K) = \widehat{P}_1(o_1) \prod_{k=2}^{K} \widehat{P}_k(o_k|\overleftarrow{o}_{k-1})$ denote the estimated distribution of $\overleftarrow{O}_K$ derived

using the above procedure. The random forest algorithm for estimating $P_k(o_k|\overleftarrow{o}_{k-1})$ involves the following steps:

(1) Create B sets of bootstrap weights $\mathbf{W}_k^{(b)} = \left\{W_{k,1}^{(b)}, \ldots, W_{k,n}^{(b)}\right\}, b = 1, \ldots, B$. Each set is independently drawn from a multinomial distribution with n "trials" and n "event probabilities", all equal to $1/n$. The bth set of weights will be used to grow the bth tree. In words, $W_{k,i}^{(b)}$ is the bootstrap weight corresponding to observation i $(i = 1, \ldots, n)$ in tree b $(b = 1, \ldots, B)$ used in the random forest algorithm associated with the kth $(k = 2, \ldots, K)$ conditional distribution.

(2) For each set of bootstrap weights, build a fully grown CART tree, where at each splitting decision only `mtry` (`mtry` $< k - 1$) randomly selected covariates are considered for splitting. From these trees, we create B estimated conditional probabilities of $O_k = o_k$ given $\overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1}$, which we denote by $\widehat{\phi}_k^{(b)}(o_k; \overleftarrow{o}_{k-1}), b = 1, \ldots, B$. The aggregated estimated conditional distribution, $\widehat{P}_k(o_k|\overleftarrow{o}_{k-1}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\phi}_k^{(b)}(o_k; \overleftarrow{o}_{k-1})$, is our estimator of $P_k(o_k|\overleftarrow{o}_{k-1})$.

The sample space of O_k is ternary. It follows that the cardinality of the sample space of $\overleftarrow{O}_k$ is finite. We refer to the distinct values of the sample space of $\overleftarrow{O}_k$ that occur with a positive probability as "histories" and denote them by $\overleftarrow{o}_k^{(1)}, \ldots, \overleftarrow{o}_k^{(L_k)}$, where $L_k \leqslant 3^k$. Using the finite cardinality of the sample space of $\overleftarrow{O}_{k-1}$ and that K, B and D are fixed, the probability that each bootstrap sample has at least D $\overleftarrow{o}_{k-1}^{(l)}$, for all $l = 1, \ldots, L_{k-1}$, converges to one as $n \to \infty$. As the random forest algorithm uses fully grown trees, it follows that asymptotically all trees in random forest algorithm k will have the same terminal nodes (with probability one) and each terminal node will correspond to a specific history $\overleftarrow{o}_{k-1}^{(l)}$.

Hence, the random forest estimator of $P_k(o_k|\overleftarrow{o}_{k-1})$ is only defined for $\overleftarrow{o}_{k-1} = \overleftarrow{o}_{k-1}^{(l)}$ and $(\widehat{P}_k((1,0)|\overleftarrow{o}_{k-1}^{(l)}), \widehat{P}_k((0,1)|\overleftarrow{o}_{k-1}^{(l)}))$ $(l = 1, \ldots, L_{k-1})$ is asymptotically equivalent to

$$(\widehat{\psi}_k((1,0)|\overleftarrow{o}_{k-1}^{(l)}), \widehat{\psi}_k((0,1)|\overleftarrow{o}_{k-1}^{(l)})),$$

where, for $o_k = (1,0),(0,1)$, $\widehat{\psi}_k(o_k|\overleftarrow{o}_{k-1}^{(l)}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\psi}_k^{(b)}(o_k|\overleftarrow{o}_{k-1}^{(l)})$ and

$$\widehat{\psi}_k^{(b)}(o_k|\overleftarrow{o}_{k-1}^{(l)}) = \frac{\sum_{i=1}^{n} W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}, O_{k,i} = o_k)}{\sum_{i=1}^{n} W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)})} \text{ for } o_k = (1,0),(0,1).$$

Note that $\widehat{\psi}_k^{(b)}(o_k|\overleftarrow{o}_{k-1}^{(l)})$ is the bth bootstrap weighted proportion of outcomes O_k that equal o_k among those with $\overleftarrow{O}_{k-1}$ equal to $\overleftarrow{o}_{k-1}^{(l)}$. Using this asymptotic equivalence, we can prove (see Appendix) that the joint distribution of $\overleftarrow{O}_K$ (with support in the set $\{\overleftarrow{o}_K^{(l)} : l = 1,\ldots,L_K\}$) is $\sqrt{n}$ consistent and asymptotic normal.

THEOREM 2: *$\widehat{P}_K(\cdot)$ is $\sqrt{n}$ consistent and asymptotically normal.*

This theorem implies that our plug-in estimator of $P(\overleftarrow{Y}_K^{(1)} = \overleftarrow{y}_K^{(1)})$ (and functionals thereof), for specified $\overleftarrow{\alpha}_K$, will be $\sqrt{n}$ consistent and asymptotically normal. The asymptotic normality implies asymptotic consistency of the empirical bootstrap estimator (Van Der Vaart and Wellner, 1996, Chapter 3.6 and 3.9.3). This provides theoretical justification for using the non-parametric bootstrap for confidence interval constructions.

5. Re-Analysis of CTN-0044

In the primary analysis of CTN-0044, individuals were defined, at each half-week, as abstinent if their urine screen (for ten drugs) was negative *and* their self-report indicated no drug use/heavy drinking days and not abstinent if their urine screen was positive even if their self-report was missing; otherwise abstinence status was treated as missing. The 24 half-weeks were analyzed using a logistic regression model, with site and primary substance (stimulant vs. non-stimulant) as main effects, a linear time-by-treatment interaction during the first 16 half-weeks and a constant treatment effect during the last 8 half-weeks (primary estimand) . Generalized estimating equations were used to account for correlation of outcomes within individuals. The validity of the inference about the primary estimand is based on the missing completely at random (MCAR) assumption (i.e., missingness is independent of outcomes,

conditional on site, primary substance and treatment group). The estimated odds ratio of abstinence during the last 8 half-weeks (TAU+ vs. TAU) was 1.62 (95% CI: 1.12 to 2.35).

For illustrative purposes, our analysis focuses on the first 6 weeks of urine data (i.e., $K = 12$). We let $Y_k^{(1)}$ be the indicator of a negative urine sample at visit k. For each treatment group, we are interested in drawing inference about the mean number of negative urine samples, i.e., $E[\sum_{k=1}^{12} Y_k^{(1)}] = \sum_{k=1}^{12} E[Y_k^{(1)}]$. Table 1 summarizes missingness patterns by treatment group. The table shows lower rates of missing data for the TAU+ arm.

[Table 1 about here.]

We first used the random forest algorithm to estimate the distribution of the observed data. We used 1000 trees. To evaluate the model fit, we compared empirical and model-based estimates of the joint distribution of the observed data at all 66 pairs of time points. For each pair, the joint distribution is represented by the cell probabilities of a three by three table. For each table, we computed the maximum of the absolute differences between the empirical and model-based estimates of the cell probabilities. The largest of these maximums over the 66 tables was 1.82%. In contrast, the largest of the maximums based on a first-order Markov model was 12.98%. This exercise demonstrates the outstanding modeling capability of the random forest algorithm.

For each treatment group, we estimated the average number of negative urine samples during the first 6 weeks under the following assumptions: missing completely at random (MCAR), missing equals positive, missing equals negative and under the benchmark assumption (i.e., (1) with $\alpha_k = 0$). Table 2 displays the treatment-specific estimates and difference in estimates, along with 95% symmetric percentile bootstrap confidence intervals (parametric bootstrap for benchmark analysis; non-parametric bootstrap for other analyses; 5000 samples). With the exception of the missing equals abstinent analysis, all analyses are consistent with treatment effects favoring TAU+, with an estimated difference in the number

abstinent days on the order of 1 day. Relative to MCAR, the treatment-specific estimates of the mean number of negative samples is lower under the benchmark assumption, indicating that the missing urine samples are more likely (under the benchmark assumption) to be positive than those observed.

[Table 2 about here.]

In our analysis, we assumed $\alpha_k = \alpha$ over all k. For each treatment group and each α (ranging from -10 to 10), we estimated the mean number of negative samples. The results (along with 95% symmetric percentile bootstrap confidence intervals - parametric bootstrap; 5000 samples) are presented in Figure 2. Notice that as $\alpha \to \infty$ and $\alpha \to -\infty$, the estimates trend toward the missing=positive and missing=negative assumptions, respectively. Figure 3 is a contour plot which shows that inferences are highly sensitive, as one deviates away from the benchmark assumption. For example, when the sensitivity analysis parameters in the TAU and TAU+ groups are 0.0 and -1.0, respectively, the estimated difference is 0.47 (95% CI: -0.39 to 1.32), indicating more equivocal evidence of a treatment relative to the benchmark analysis.

[Figure 2 about here.]

[Figure 3 about here.]

To better understand the choice of sensitivity analysis parameters, consider Figure 4. For each treatment group, we display as a function of α, the percent difference between the estimated probability of a negative urine sample at visit k ($k = 1, \ldots, 12$) for those who do not provide a sample and the observed proportion of negative urine samples among those who provide a sample. At $\alpha = 0$, the percent difference ranges across visits from -41% and -14% for TAU and from -42% and -19% for TAU+. At $\alpha = 10$ ($\alpha = -10$) these ranges are 36% to 77% (-100% to -79%) and 24% to 48% (-100% to -94%) for TAU and TAU+, respectively.

While it is impossible to determine the true value of α, substance abuse researchers can use their scientific and clinical judgment to rule out specific choices. Clinical experience suggests that when patients with a substance use disorder avoid appointments, they are likely not doing well and using substances. It is therefore reasonable to assume that, ceteris paribus, individuals in treatment for substance use disorders, who miss a visit and do not provide urine samples, have a lower chance of testing negative than individuals who provide a urine sample. Thus, $\alpha > 0$ can be ruled out. By focusing on the lower left quadrant of Figure 3, we see there is great sensitivity of inferences.

[Figure 4 about here.]

6. Simulation Study

We used the treatment-specific estimates of the distribution of the observed data computed using the random forest algorithm as the true observed data generating mechanism. We evaluate the performance of our procedures for various $\alpha_k = \alpha$ values ranging from -5 to 5. For each α, the true treatment-specific mean number of negative visits was computed using Theorem 1. In the simulation, the sample size for each dataset was 250 and the number of datasets generated was set to 500. For the computation of symmetric, percentile parametric bootstrap confidence intervals, 1000 samples were generated. The random forest algorithm applied to the generated datasets used 500 trees. The results of the simulation are shown in Table 3. The table shows low bias and coverage close to the nominal level.

[Table 3 about here.]

7. Software and Implementation

The methods have been implemented in the R package `salbm`. The package can be installed in `R` from Github by `install_github("olssol/salbm")`. A web-application for the method has also been developed using `R Shiny` and is included in `salbm`.

8. Discussion

In this paper, we developed a novel global sensitivity analysis procedure for the analysis randomized trials in which (1) participants are scheduled to have binary outcomes assessed at fixed points in time after randomization and (2) some of the outcomes may be missing in a non-monotone or intermittent fashion. Our procedure was built using random forests to model the distribution of the observed data. We established $\sqrt{n}$ asymptotic theory for the random forest estimator of the distribution of the observed data and, using the plug-in principle, established $\sqrt{n}$ asymptotics for smooth functionals.

There is a key computational limitation to our approach. It requires storage and operation on a 3^K vector of probabilities. This starts to become computationally infeasible when $K > 15$. To address this problem, we plan, in a follow-up manuscript, to reduce the dimension of our model by introducing Markovian-type conditional independence restrictions.

Another next step is to develop an extension to handle continuous outcomes. Unfortunately, the asymptotic theory for the random forest estimator is substantially more complex for continuous outcomes. Previous work establishing asymptotic normality of predictions from the random forest algorithm in the continuous case has deviated substantially from the traditional random forest algorithm, e.g., simplifying the theoretical developments by using subsampling instead of bootstrapping. Furthermore, either no rate of convergence results are provided (Mentch and Hooker, 2016) or strict assumptions are imposed on the trees that serve as building blocks of the forests (Wager and Athey, 2018). Thus, the continuous case

will likely require a different strategy, e.g., using an influence function-based approach as in Scharfstein et al. (2018).

Appendix

LEMMA 3: *For a fixed L, let $A_n, B_{n,1}, \ldots, B_{n,L}$ be $p_0 \times 1, \ldots, p_L \times 1$ random vectors and constants $t_l \in \mathbb{R}^{p_l}, l = 0, \ldots, L$ be given. Let R_n be a sequence of random quantities. Denote $\psi_n(t_0) = E[\exp\{it_0^T A_n\}]$ and $\gamma_{n,l}(t_l|R_n) = E[\exp\{it_l^T B_{n,l}\}|R_n], l = 1, \ldots, L$, as the characteristic functions associated with the marginal distribution of A_n and the conditional distributions of $B_{n,l}$ given R_n, respectively. Let A and $B_l, l = 1, \ldots, L$, be random vectors with characteristic functions $\psi(t_0) = E[\exp\{it_0^T A\}]$ and $\gamma_l(t_l) = E[\exp\{it_l^T B_l\}]$, respectively. Assume the following:*

A.1 The random variable A_n is a deterministic function of R_n.

A.2 For $l \neq l'$, $B_{n,l}$ and $B_{n,l'}$ are independent conditioned on R_n.

A.3 The sequence of characteristic functions $\psi_n(t_0) \to \psi(t_0)$ and $\psi(t_0)$ is a deterministic function of t_0.

A.4 For $l = 1, \ldots, L$, the sequence of characteristic functions $\gamma_l(t_l|R_n) \to \gamma_l(t_l)$ in probability and $\gamma_l(t_l)$ is a deterministic function of t_l.

Under Assumptions $A.1 - A.4$,

$$(A_n^T, B_{n,1}^T, \ldots, B_{n,L}^T)^T \text{ converges in distribution to } (A^T, B_1^T, \ldots, B_L^T)^T$$

and the components of the limiting vector are independent.

Proof. For $l \in \{1, \ldots, L\}$, let $\mathbf{t}_l = (t_0^T, \ldots, t_l^T)^T$ and define

$$\alpha_{n,l}(\mathbf{t}_l) = E[\exp\{i(\mathbf{t}_l)^T (A_n^T, B_{n,1}^T \ldots, B_{n,l}^T)^T\}]$$

as the characteristic function of $(A_n^T, B_{n,1}^T, \ldots, B_{n,l}^T)^T$, and

$$\alpha_l(\mathbf{t}_l) = E[\exp\{i(\mathbf{t}_l)^T (A^T, B_1^T \ldots, B_l^T)^T\}]$$

as the characteristic function of $(A^T, B_1^T, \ldots, B_l^T)^T$.

The proof uses induction. Start by proving the result when $L = 1$. Using Assumption $A.1$ and the tower rule for conditional expectations,

$$\begin{aligned}
\alpha_{n,1}(\mathbf{t}_1) &= E\left[\exp\{it_0^T A_n\}\exp\{it_1^T B_{n,1}\}\right] \\
&= E\left[\exp\{it_0^T A_n\}E[\exp\{it_1^T B_{n,1}\}|R_n]\right] \\
&= E\left[\exp\{it_0^T A_n\}\gamma_{n,1}(t_1|R_n)\right] \\
&= E\left[\exp\{it_0^T A_n\}\left\{\gamma_{n,1}(t_1|R_n) - \gamma_1(t_1)\right\}\right] + \left\{\psi_n(t_0) - \psi(t_0)\right\}\gamma_1(t_1) + \psi(t_0)\gamma_1(t_1)
\end{aligned}$$

Hence,

$$|\alpha_{n,1}(\mathbf{t}_1) - \psi(t_0)\gamma_1(t_1)| \leqslant |\psi_n(t_0) - \psi(t_0)||\gamma_1(t_1)| + E[|\exp\{it_0^T A_n\}||\gamma_{n,1}(t_1|R_n) - \gamma_1(t_1)|]$$

As $|\gamma_1(t_1)| = |E[\exp\{it_1^T B_1\}]| \leqslant 1$ and $|\exp\{it_0^T A_n\}| \leqslant 1$, it follows from Assumptions $A.3$ and $A.4$ that

$$|\alpha_{n,1}(\mathbf{t}_1) - \psi(t_0)\gamma_1(t_1)| \to 0$$

The result for $L = 1$ follows from Levy's continuity theorem as $\psi(t_0)\gamma_1(t_1)$ is the characteristic function of $(A^T, B_1^T)^T$ with A and B_1 independent.

Now assume that $(A_n^T, B_{n,1}^T, \ldots, B_{n,L-1}^T)$ converges in distribution to $(A^T, B_1^T, \ldots, B_{L-1}^T)$ and the components of the limiting vector are independent. We want to show that $(A_n^T, B_{n,1}^T, \ldots, B_{n,L}^T)$ converges in distribution to $(A^T, B_1^T, \ldots, B_L^T)$ and the components of the limiting vector are independent.

By the induction assumption

$$\alpha_{n,L-1}(\mathbf{t}_{L-1}) \to \alpha_{L-1}(\mathbf{t}_{L-1}) = \psi(t_0)\prod_{l=1}^{L-1}\gamma_l(t_l).$$

Using the tower rule for conditional expectations and Assumptions $A.1 - A.4$ gives

$$\begin{aligned}
&\alpha_{n,L}(\mathbf{t}_L) \\
&= E\left[E\left[\exp\{i(t_0^T,\ldots,t_L^T)(A_n,B_{n,1}^T,\ldots,B_{n,L}^T)^T\}|R_n\right]\right] \\
&= E\left[\exp\{it_0^TA_n\}E\left[\exp\{i(t_1^T,\ldots,t_L^T)(B_{n,1}^T,\ldots,B_{n,L}^T)^T\}|R_n\right]\right] \\
&= E\left[\exp\{it_0^TA_n\}E[\exp\{it_L^TB_{n,L}\}|R_n]E\left[\exp\{i(t_1^T,\ldots,t_{L-1}^T)(B_{n,1}^T,\ldots,B_{n,L-1}^T)^T\}|R_n\right]\right] \\
&= E\left[E\left[\exp\{i(t_0^T,t_1^T,\ldots,t_{L-1}^T)(A_n^T,B_{n,1}^T,\ldots,B_{n,L-1}^T)^T\}E[\exp\{it_L^TB_{n,L}\}|R_n]|R_n\right]\right] \\
&= E\left[\exp\{i(t_0^T,t_1^T,\ldots,t_{L-1}^T)(A_n^T,B_{n,1}^T,\ldots,B_{n,L-1}^T)^T\}E[\exp\{it_L^TB_{n,L}\}|R_n]\right] \\
&= E\left[\exp\{i(t_0^T,t_1^T,\ldots,t_{L-1}^T)(A_n^T,B_{n,1}^T,\ldots,B_{n,L-1}^T)^T\}\gamma_{n,L}(t_L|R_n)\right] \\
&= E\left[\exp\{i(t_0^T,t_1^T,\ldots,t_{L-1}^T)(A_n^T,B_{n,1}^T,\ldots,B_{n,L-1}^T)^T\}\{\gamma_{n,L}(t_L|R_n)-\gamma_L(t_L)\}\right]+ \\
&\qquad \{\alpha_{n,L-1}(\mathbf{t}_{L-1})-\alpha_{L-1}(\mathbf{t}_{L-1})\}\gamma_L(t_L)+\alpha_{L-1}(\mathbf{t}_{L-1})\gamma_L(t_L)
\end{aligned}$$

Using the calculation above and the induction hypothesis

$$|\alpha_{n,L}(\mathbf{t}_L)-\psi(t_0)\prod_{l=1}^{L}\gamma_l(t_l)| \leqslant |\alpha_{n,L-1}(\mathbf{t}_{L-1})-\alpha_{L-1}(\mathbf{t}_{L-1})||\gamma_L(t_L)|+$$
$$E\left[|\exp\{i(\mathbf{t}_{L-1})^T(A_n^T,B_{n,1}^T\ldots,B_{n,L-1}^T)^T\}||\gamma_{n,L}(t_L|R_n)-\gamma_L(t_L)|\right]$$

As $|\gamma_L(t_L)| \leqslant 1$ and $|\exp\{i(\mathbf{t}_{L-1})^T(A_n^T,B_{n,1}^T\ldots,B_{n,L-1}^T)^T\}| \leqslant 1$, it follows by the induction hypothesis and Assumption $A.4$ that

$$\alpha_{n,L}(\mathbf{t}_L) \to \psi(t_0)\prod_{l=1}^{L}\gamma_l(t_l).$$

The function $\psi(t_0)\prod_{l=1}^{L}\gamma_l(t_l)$ is the characteristic function of $(A^T,B_1^T,\ldots,B_L^T)^T$ with all components of the vector being independent. It follows from Levy's continuity theorem that $(A_n^T,B_{n,1}^T,\ldots,B_{n,L}^T)^T$ converges in distribution to $(A^T,B_1^T,\ldots,B_L^T)^T$.

For a function $f_k:\Omega_k \to \mathbb{R}^{q_k}, k=1,\ldots,K$, where Ω_k is the sample space for $\overleftarrow{O}_k$, and for a given $b \in \{1,\ldots,B\}$, define

$$\mathbb{P}_{n,k}^{(b)}(f_k) = \frac{1}{n}\sum_{i=1}^{n} W_{k,i}^{(b)} f_k(\overleftarrow{O}_{k,i})^T,$$

$$\mathbb{P}_{n,k}(f_k) = \frac{1}{n}\sum_{i=1}^{n} f_k(\overleftarrow{O}_{k,i})^T,$$

and

$$\mathbb{P}_k(f_k) = E[f_k(\overleftarrow{O}_k)^T].$$

LEMMA 4: *If, for all $k \in \{1,\ldots,K\}$, $f_k : \Omega_k \to \mathbb{R}^{q_k}$ has finite covariance matrix, then*

$$\sqrt{n}\left(\mathbb{P}_{n,1}(f_1) - \mathbb{P}_1(f_1), \frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,2}^{(b)}(f_2) - \mathbb{P}_2(f_2), \ldots, \frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,K}^{(b)}(f_K) - \mathbb{P}_K(f_K)\right)^T$$

is asymptotically normal.

Proof. For compactness of notation throughout the proof of the lemma we drop the dependence on f_k when writing $\mathbb{P}_{n,k}^{(b)}(f_k)$, $\mathbb{P}_{n,k}(f_k)$, and $\mathbb{P}_k(f_k)$.

We start by showing that the conditions of Lemma 3 hold when

$$A_n = \sqrt{n}((\mathbb{P}_{n,1} - \mathbb{P}_1)^T, \ldots, (\mathbb{P}_{n,K} - \mathbb{P}_K)^T)^T \tag{A.1}$$

$$B_{n,l} = \sqrt{n}\left(\mathbb{P}_{n,k}^{(b)} - \mathbb{P}_{n,k}\right), \quad l \in \{(b,k) : k = 2,\ldots,K, b = 1,\ldots,B\}, \tag{A.2}$$

and R_n is the observed data $\mathcal{O}_n$.

As $\sqrt{n}((\mathbb{P}_{n,1}-\mathbb{P}_1)^T, \ldots, (\mathbb{P}_{n,K}-\mathbb{P}_K)^T)^T$ is a deterministic function of $\mathcal{O}_n$, Assumption $A.1$ in Lemma 3 is satisfied. Assumption $A.3$ of Lemma 3 follows since $\sqrt{n}((\mathbb{P}_{n,1}-\mathbb{P}_1)^T, \ldots, (\mathbb{P}_{n,K}-\mathbb{P}_K)^T)^T$ is asymptotically normal by the central limit theorem.

By the independence of the bootstrap weights, $W_{k,i}^{(b)}$ is independent of $W_{k',i'}^{(b')}$ for $(b,k) \neq (b',k'), b \in \{1,\ldots,B\}, k \in \{2,\ldots,K\}, i, i' \in \{1,\ldots,n\}$. Hence for $(b,k) \neq (b',k')$, $\sqrt{n}\left(\mathbb{P}_{n,k}^{(b)} - \mathbb{P}_{n,k}\right)$ is independent of $\sqrt{n}\left(\mathbb{P}_{n,k'}^{(b')} - \mathbb{P}_{n,k'}\right)$ conditioned on the observed data $\mathcal{O}_n$. This implies that Assumption $A.2$ from Lemma 3 is satisfied.

Example 3.6.10 in Van Der Vaart and Wellner (1996) shows that the multinomial bootstrap weights satisfy the conditions required for Theorem 3.6.13 in Van Der Vaart and Wellner (1996) to hold. This theorem combined with the Portmanteau Theorem imply that, $\sqrt{n}\left(\mathbb{P}_{n,k}^{(b)} - \mathbb{P}_{n,k}\right)$ is asymptotically normal conditioned on the observed data $\mathcal{O}_n$ for all pairs

$(b,k), k \in \{2,\ldots,K\}, b \in \{1,\ldots,B\}$ and the limit is the same for almost all sequences of data. It follows that Assumption $A.4$ of Lemma 3 is satisfied.

Lemma 3 gives that when A_n and $B_{n,l}$ are defined as in (A.1) and (A.2),

$$\sqrt{n}\Bigg((\mathbb{P}_{n,1}-\mathbb{P}_1)^T,\ldots,(\mathbb{P}_{n,K}-\mathbb{P}_K)^T,\left(\mathbb{P}_{n,2}^{(1)}-\mathbb{P}_{n,2}\right)^T,\ldots,\left(\mathbb{P}_{n,2}^{(B)}-\mathbb{P}_{n,2}\right)^T,$$
$$\ldots,\left(\mathbb{P}_{n,K}^{(1)}-\mathbb{P}_{n,K}\right)^T,\ldots,\left(\mathbb{P}_{n,K}^{(B)}-\mathbb{P}_{n,K}\right)^T\Bigg)^T$$

has a limiting distribution which is normally distributed.

We can write

$$\sqrt{n}\left(\left(\mathbb{P}_{n,1}-\mathbb{P}_1\right)^T,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,2}^{(b)}-\mathbb{P}_2\right)^T,\ldots,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,K}^{(b)}-\mathbb{P}_K\right)^T\right)^T$$
$$=\sqrt{n}\left(\left(\mathbb{P}_{n,1}-\mathbb{P}_1\right)^T,\ldots,\left(\mathbb{P}_{n,K}-\mathbb{P}_K\right)^T\right)^T$$
$$+\sqrt{n}\left(\mathbf{0},\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,2}^{(b)}-\mathbb{P}_{n,2}\right)^T,\ldots,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,K}^{(b)}-\mathbb{P}_{n,K}\right)^T\right)^T.$$

Using the decomposition above,

$$\sqrt{n}\left(\left(\mathbb{P}_{n,1}-\mathbb{P}_1\right)^T,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,2}^{(b)}-\mathbb{P}_2\right)^T,\ldots,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,K}^{(b)}-\mathbb{P}_K\right)^T\right)^T$$

is a linear function of

$$\sqrt{n}\Bigg((\mathbb{P}_{n,1}-\mathbb{P}_1)^T,\ldots,(\mathbb{P}_{n,K}-\mathbb{P}_K)^T,\left(\mathbb{P}_{n,2}^{(1)}-\mathbb{P}_{n,2}\right)^T,\ldots,\left(\mathbb{P}_{n,2}^{(B)}-\mathbb{P}_{n,2}\right)^T,$$
$$\ldots,\left(\mathbb{P}_{n,K}^{(1)}-\mathbb{P}_{n,K}\right)^T,\ldots,\left(\mathbb{P}_{n,K}^{(B)}-\mathbb{P}_{n,K}\right)^T\Bigg)^T$$

It follows that,

$$\sqrt{n}\left(\left(\mathbb{P}_{n,1}-\mathbb{P}_1\right)^T,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,2}^{(b)}-\mathbb{P}_2\right)^T,\ldots,\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,K}^{(b)}-\mathbb{P}_K\right)^T\right)^T$$

is asymptotically normal.

Proof of Theorem 2

Proof. Consistency follows from the asymptotic equivalence of $\widehat{\psi}_k(\cdot|\overleftarrow{o}_{k-1}^{(l)})$ and $\widehat{P}_k(\cdot|\overleftarrow{o}_{k-1}^{(l)})$, the bootstrap consistency results of Csörgo (1992) and the continuous mapping theorem.

Set $f_1^{(1)}(\overleftarrow{O}_1) = (I(O_1 = (1,0)), I(O_1 = (0,1)))^T$ and, for $k = 2, \ldots, K$,

$$f_k^{(1)}(\overleftarrow{O}_k | \overleftarrow{o}_{k-1}^{(l)}) = \frac{I(\overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1}^{(l)}) \left(I(O_k = (1,0)), I(O_k = (0,1))\right)^T}{P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}.$$

Note that

$$\mathbb{P}_k(f_k^{(1)}(\overleftarrow{O}_k | \overleftarrow{o}_{k-1}^{(l)})) = (P_k((1,0) | \overleftarrow{o}_{k-1}^{(l)}), P_k((0,1) | \overleftarrow{o}_{k-1}^{(l)}))^T$$

With this choice of $f_k^{(1)}(\overleftarrow{O}_k | \overleftarrow{o}_{k-1}^{(l)})$ $(k = 2, \ldots, K)$, notice that

$$\mathbb{P}_{n,k}^{(b)}(f_k^{(1)}) = \frac{\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) \left(I(O_{k,i} = (1,0)), I(O_{k,i} = (0,1))\right)}{n P_{k-1}(\overleftarrow{o}_{k-1}^{(l)}))}$$

which is equivalent to $(\widehat{\psi}_k^{(b)}((1,0) | \overleftarrow{o}_{k-1}^{(l)}), \widehat{\psi}_k^{(b)}((0,1) | \overleftarrow{o}_{k-1}^{(l)}))$ except that $\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)})$ is replaced by $n P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})$.

Now,

$$\begin{aligned}
&\left(\left(\widehat{\psi}_k\left((1,0) | \overleftarrow{o}_{k-1}^{(l)}\right), \widehat{\psi}_k\left((0,1) | \overleftarrow{o}_{k-1}^{(l)}\right)\right) - \mathbb{P}_k(f_k^{(1)})\right) \\
&= \left(\frac{1}{B}\sum_{b=1}^B \mathbb{P}_{n,k}^{(b)}(f_k^{(1)}) - \mathbb{P}_k(f_k^{(1)})\right) - \\
&\quad \left(\frac{1}{B}\sum_{b=1}^B \left\{ \frac{\frac{1}{n}\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) \left(I(O_{k,i} = (1,0)), I(O_{k,i} = (0,1))\right)}{\frac{1}{n}\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})} \right\} \times \right. \\
&\qquad \left. \left\{ \frac{1}{n}\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) - P_{k-1}(\overleftarrow{o}_{k-1}^{(l)}) \right\} \right)
\end{aligned} \tag{A.3}$$

By bootstrap consistency (Csörgo, 1992),

$$\begin{aligned}
&\frac{\frac{1}{n}\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) \left(I(O_{k,i} = (1,0)), I(O_{k,i} = (0,1))\right)}{\frac{1}{n}\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})} - \\
&\qquad \frac{E[\left(I(O_k = (1,0)), I(O_k = (0,1))\right) | \overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1}^{(l)}]}{P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})} = o_P(1).
\end{aligned} \tag{A.4}$$

It follows from the proof of Lemma 4 that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n W_{k,i}^{(b)} I(\overleftarrow{O}_{k-1,i} = \overleftarrow{o}_{k-1}^{(l)}) - P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})\right) = O_P(1). \tag{A.5}$$

Combining equations (A.4) and (A.5) and Slutky's theorem, we can write the second term

on the right hand side of (A.3) as

$$\begin{aligned}
&\left(\frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})\left(I(O_{k,i}=(1,0)),I(O_{k,i}=(0,1))\right)}{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}\right\}\times\right.\\
&\qquad\left.\left\{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})-P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})\right\}\right)\\
&=\left(\frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})\left(I(O_{k,i}=(1,0)),I(O_{k,i}=(0,1))\right)}{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}\right.\right.\\
&\qquad\left.\left.-\frac{E[(I(O_k=(1,0)),I(O_k=(0,1)))\,|\,\overleftarrow{O}_{k-1}=\overleftarrow{o}_{k-1}^{(l)}]}{P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}\right\}\times\right.\\
&\qquad\left.\left\{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})-P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})\right\}\right)+\\
&\quad\left(\frac{1}{B}\sum_{b=1}^{B}\frac{E[(I(O_k=(1,0)),I(O_k=(0,1)))\,|\,\overleftarrow{O}_{k-1}=\overleftarrow{o}_{k-1}^{(l)}]}{P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}\times\right.\\
&\qquad\left.\left\{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})-P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})\right\}\right)\\
&=\left(\frac{1}{B}\sum_{b=1}^{B}\frac{E[(I(O_k=(1,0)),I(O_k=(0,1)))\,|\,\overleftarrow{O}_{k-1}=\overleftarrow{o}_{k-1}^{(l)}]}{P(\overleftarrow{O}_{k-1}=\overleftarrow{o}_{k-1}^{(l)})}\times\right.\\
&\qquad\left.\left\{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})-P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})\right\}\right)+o_P(1/\sqrt{n}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\left(\left(\widehat{\psi}_k\left((1,0)|\overleftarrow{o}_{k-1}^{(l)}\right),\widehat{\psi}_k\left((0,1)|\overleftarrow{o}_{k-1}^{(l)}\right)\right)-\mathbb{P}_k(f_k^{(1)})\right)\\
&=\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,k}^{(b)}(f_k^{(1)})-\mathbb{P}_k(f_1^{(1)})\right)-\\
&\quad\left(\frac{1}{B}\sum_{b=1}^{B}\frac{E[(I(O_k=(1,0)),I(O_k=(0,1)))\,|\,\overleftarrow{O}_{k-1}=\overleftarrow{o}_{k-1}^{(l)}]}{P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}\times\right.\\
&\qquad\left.\left\{\frac{1}{n}\sum_{i=1}^{n}W_{k,i}^{(b)}I(\overleftarrow{O}_{k-1,i}=\overleftarrow{o}_{k-1}^{(l)})-P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})\right\}\right)+o_P(1/\sqrt{n})\\
&=\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,k}^{(b)}(f_k^{(2)})-\mathbb{P}_k(f_k^{(2)})\right)+o_P(1/\sqrt{n}),
\end{aligned}\tag{A.6}$$

where

$$f_k^{(2)}(\overleftarrow{O}_k|\overleftarrow{o}_{k-1}^{(l)})$$
$$= f_k^{(1)}(\overleftarrow{O}_k|\overleftarrow{o}_{k-1}^{(l)}) - I(\overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1}^{(l)})\frac{E[(I(O_k=(1,0)), I(O_k=(0,1)))\,|\overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1}^{(l)}]}{P_{k-1}(\overleftarrow{o}_{k-1}^{(l)})}.$$

In the second term of $f_k^{(2)}(\overleftarrow{O}_k|\overleftarrow{o}_{k-1}^{(l)})$, only $I(\overleftarrow{O}_{k-1} = \overleftarrow{o}_{k-1}^{(l)})$ is summed over (indexed by i) in the definition of $\mathbb{P}_{n,k}^{(b)}(f_k^{(2)})$ and $\mathbb{P}_{n,k}(f_k^{(2)})$.

Hence, for $k = 2, \ldots, K$ and all $\overleftarrow{o}_{k-1}^{(l)}$, and using the asymptotic equivalence of $\widehat{\psi}_k(\cdot|\overleftarrow{o}_{k-1}^{(l)})$ and $\widehat{P}_k(\cdot|\overleftarrow{o}_{k-1}^{(l)})$, gives

$$\{(\widehat{P}_k((1,0)|\overleftarrow{o}_{k-1}^{(l)}), \widehat{P}_k((0,1)|\overleftarrow{o}_{k-1}^{(l)})) - (P_k((1,0)|\overleftarrow{o}_{k-1}^{(l)}), P_k((0,1)|\overleftarrow{o}_{k-1}^{(l)}))\}^T$$
$$= \left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}_{n,k}^{(b)}(f_k^{(2)}) - \mathbb{P}_k(f_k^{(2)})\right) + o_P(1/\sqrt{n})$$

and

$$\{(\widehat{P}_1((1,0)), \widehat{P}_1((0,1)) - (P_1((1,0)), P_1((0,1))\}^T = \left(\mathbb{P}_n(f_1^{(1)}) - \mathbb{P}_1(f_1^{(1)})\right)$$

As $f_1^{(1)}$ and $f_k^{(2)}, k = 2, \ldots, K$, have a finite covariance matrix, Lemma 4 applies. It then follows that

$$\sqrt{n}\Big(\widehat{P}_1((1,0)) - P_1((1,0)), \widehat{P}_1((0,1)) - P_1((0,1)),$$
$$\widehat{P}_2((1,0)|\overleftarrow{o}_1^{(1)}) - P_2((1,0)|\overleftarrow{o}_1^{(1)}), \widehat{P}_2((0,1)|\overleftarrow{o}_1^{(1)}) - P_2((0,1)|\overleftarrow{o}_1^{(1)})$$
$$\ldots$$
$$\widehat{P}_2((1,0)|\overleftarrow{o}_1^{(L_1)}) - P_2((1,0)|\overleftarrow{o}_1^{(L_1)}), \widehat{P}_2((0,1)|\overleftarrow{o}_1^{(L_1)}) - P_2((0,1)|\overleftarrow{o}_1^{(L_1)})$$
$$\ldots$$
$$\widehat{P}_K((1,0)|\overleftarrow{o}_{K-1}^{(1)}) - P_K((1,0)|\overleftarrow{o}_{K-1}^{(1)}), \widehat{P}_K((0,1)|\overleftarrow{o}_{K-1}^{(1)}) - P_K((0,1)|\overleftarrow{o}_{K-1}^{(1)})$$
$$\ldots$$
$$\widehat{P}_K((1,0)|\overleftarrow{o}_{K-1}^{(L_{K-1})}) - P_K((1,0)|\overleftarrow{o}_{K-1}^{(L_{K-1})}), \widehat{P}_K((0,1)|\overleftarrow{o}_{K-1}^{(L_{K-1})}) - P_K((0,1)|\overleftarrow{o}_{K-1}^{(L_{K-1})})\Big)$$

is asymptotically normal. Using the delta method, it then follows that

$$\sqrt{n}(\widehat{P}_K(\overleftarrow{o}_K^{(1)}) - P_K(\overleftarrow{o}_K^{(1)}), \ldots, \widehat{P}_K(\overleftarrow{o}_K^{(L_K)}) - P_K(\overleftarrow{o}_K^{(L_K)}))$$

is asymptotically normal.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24,** 123–140.

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Campbell, A. N., Nunes, E. V., Matthews, A. G., Stitzer, M., Miele, G. M., Polsky, D., Turrigiano, E., Walters, S., McClure, E. A., Kyle, T. L., et al. (2014). Internet-delivered treatment for substance abuse: A multisite randomized controlled trial. *American Journal of Psychiatry* **171,** 683–690.

Cox, D. and Barndorff-Nielsen, O. (1994). *Inference and Asymptotics*, volume 52. CRC Press.

Csörgo, S. (1992). On the law of large numbers for the bootstrap mean. *Statistics & Probability Letters* **14,** 1–7.

Fitzmaurice, G. M., Lipsitz, S. R., and Weiss, R. D. (2018). Sensitivity analysis for non-monotone missing binary data in longitudinal studies: Application to the nida collaborative cocaine treatment study. *Statistical Methods in Medical Research* **28,** 3057–3073.

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

Lauritzen, S. L. and Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 321–348.

Linero, A. R. and Daniels, M. J. (2018). Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science* **33,** 198–213.

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. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88,** 125–134.

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

McPherson, S., Barbosa-Leiker, C., Burns, G. L., Howell, D., and Roll, J. (2012). Missing data in substance abuse treatment research: Current methods and modern approaches. *Experimental and Clinical Psychopharmacology* **20,** 243–250.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research* **17,** 1–41.

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

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

Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modeling for incomplete multivariate data. *Biometrika* **104,** 207–220.

Sadinle, M. and Reiter, J. P. (2018). Sequential identification of nonignorable missing data mechanisms. *Statistica Sinica* **28,** 1741–1759.

Scharfstein, D., McDermott, A., Díaz, I., Carone, M., Lunardon, N., and Turkoz, I. (2018). Global sensitivity analysis for repeated measures studies with informative drop-out: A semi-parametric approach. *Biometrics* **74,** 207–219.

Shpitser, I. (2016). Consistent estimation of functions of data nissing non-monotonically and

not at random. In *Advances in Neural Information Processing Systems*, pages 3144–3152.

Tchetgen-Tchetgen, E. J., Wang, L., and Sun, B. (2017). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica* **28,** 2069–2088.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer.

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94,** 841–860.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113,** 1228–1242.

Yang, X. and Shoptaw, S. (2005). Assessing missing data assumptions in longitudinal studies: An example using a smoking cessation trial. *Drug and Alcohol Dependence* **77,** 213–225.

Zhou, Y., Little, R. J., Kalbfleisch, J. D., et al. (2010). Block-conditional missing at random models for missing data. *Statistical Science* **25,** 517–532.

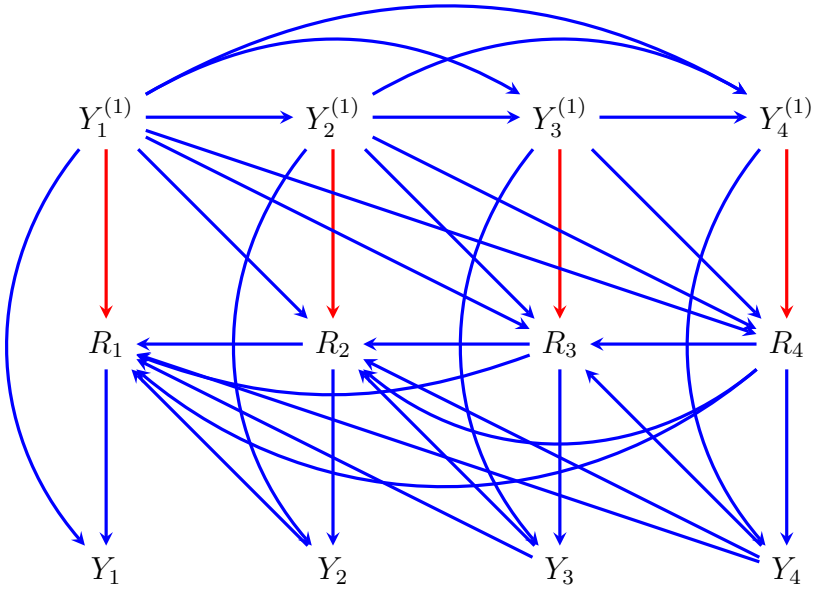
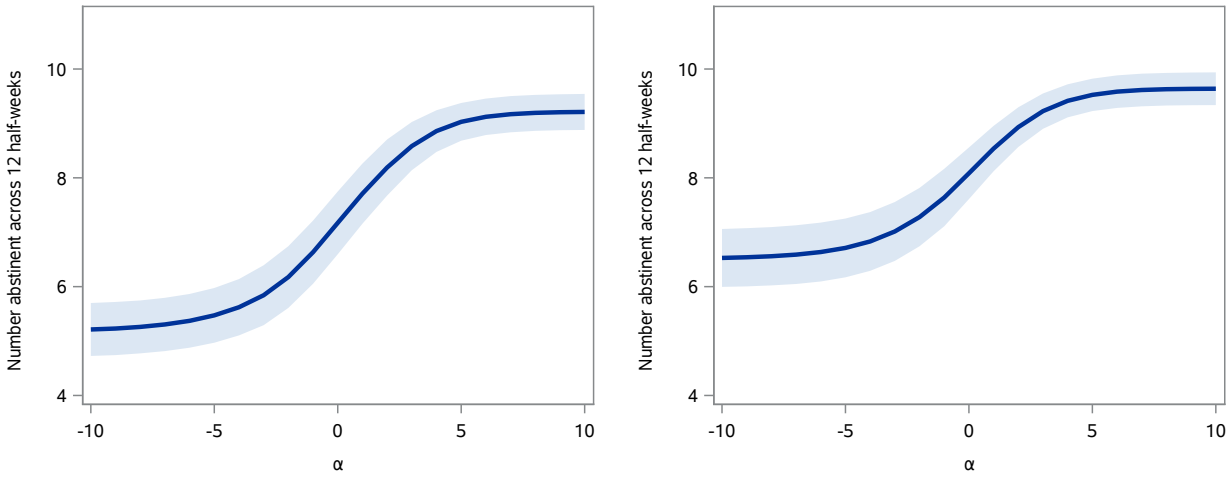


Figure 1: Directed acyclic graph representation (DAG) of (1) with $K = 4$. In this DAG, there are arrows into Y_k from $Y_k^{(1)}$ and R_k since Y_k is a deterministic function of these latter variables. The red arrow from $Y_k^{(1)}$ into R_k represents the dependence implied by the sensitivity analysis parameter α_k. The red arrows will be absent when $\alpha_k = 0$ for all k.



(a) TAU (b) TAU+

Figure 2: Treatment-specific estimates of mean number of negative urine samples (along with 95% confidence intervals) as a function of α.

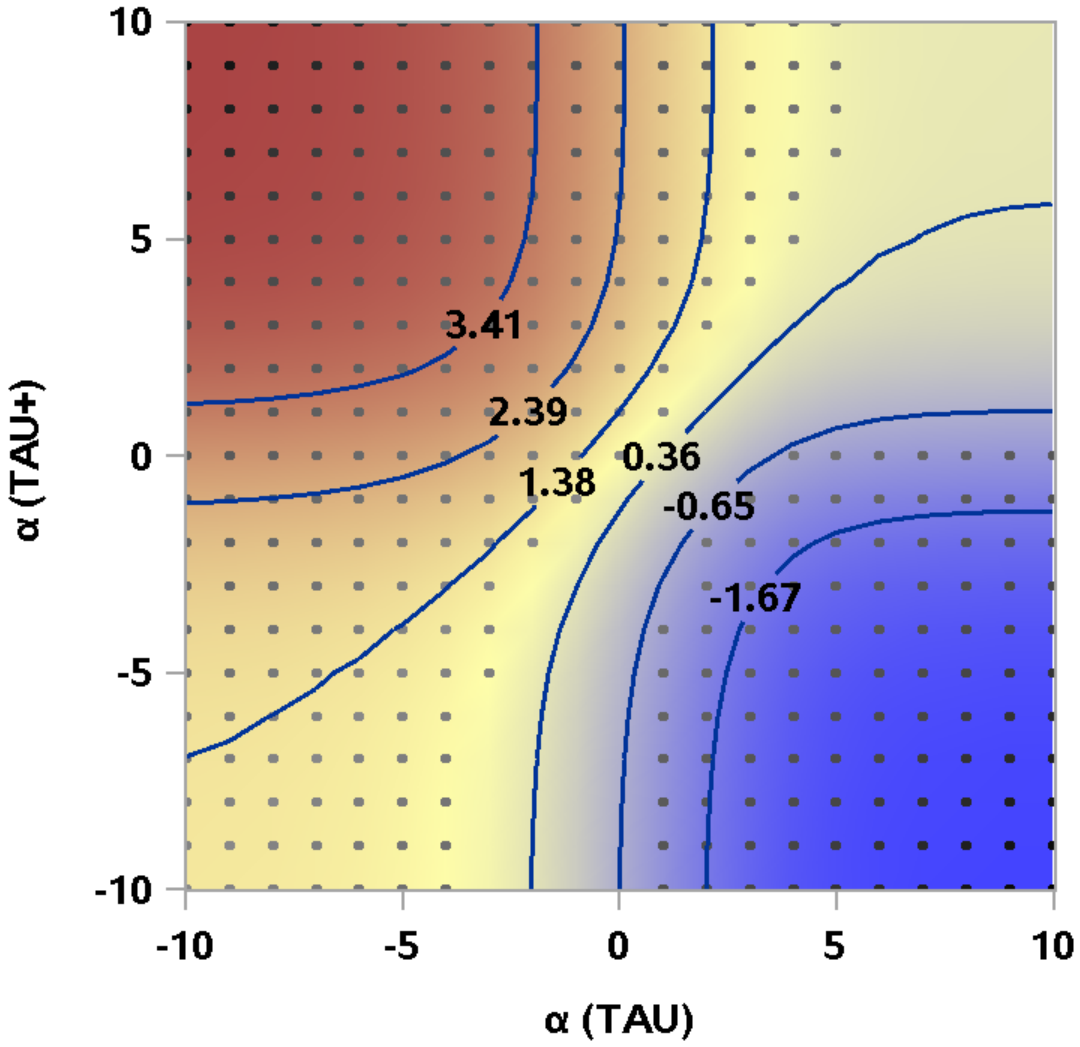
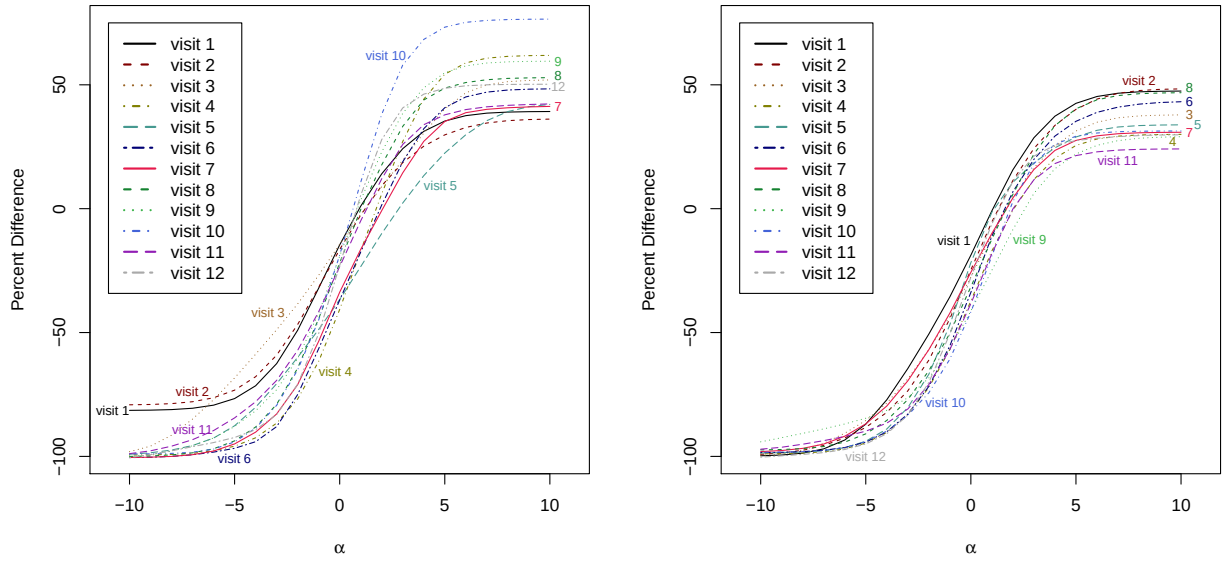


Figure 3: Contour plot of estimated treatment differences, as a function of treatment-specific sensitivity analysis parameters. Combinations of treatment-specific sensitivity parameters with a dot indicate that the associate 95% confidence interval excludes 0. Positive effects favor TAU+ and negative effects favor TAU.



(a) TAU (b) TAU+

Figure 4: Treatment- and visit- specific estimates of percent difference between the induced probability of a negative urine sample for individuals who do not provide a sample and the proportion negative among those who do provide a sample, as a function of α.

Table 1: CTN-0044: Missingness Patterns

Missingness Pattern	TAU ($n = 252$)	TAU+ ($n = 255$)
Complete	42 (16.7%)	81 (31.8%)
Monotone		
$1 - 11$ Missing	28 (11.1%)	18 (7.1%)
All Missing	11 (4.4%)	3 (1.2%)
Non-monotone		
1 Missing	35 (13.9%)	36 (14.1%)
2 Missing	32 (12.7%)	32 (12.5%)
$\geqslant 3$ Missing	104 (41.3%)	85 (33.3%)

Table 2: Inference under missing completely at random (MCAR), missing equals positive and missing equals negative as well as under the benchmark assumption (i.e., (1) with $\alpha_k = 0$)

Assumption	TAU	TAU+	Difference
MCAR	7.86 (7.25, 8.47)	8.83 (8.28, 9.38)	0.97 (0.17, 1.76).
Missing=Positive	5.14 (4.60, 5.69)	6.48 (5.90, 7.06)	1.34 (0.58, 2.10)
Missing=Negative	9.27 (8.87, 9.67)	9.64 (9.24, 10.04)	0.37 (-0.18, 0.92)
Benchmark	7.17 (6.60, 7.75)	8.08 (7.61, 8.56)	0.91 (0.06, 1.76)

Table 3: Results of simulation study

	TAU					TAU+				
α	Truth	Mean	Std. Dev.	$\sqrt{\text{MSE}}$	Coverage	Truth	Mean	Std. Dev.	$\sqrt{\text{MSE}}$	Coverage
-5	5.47	5.51	0.282	0.286	0.954	6.71	6.78	0.306	0.313	0.942
-4	5.62	5.67	0.285	0.289	0.952	6.82	6.90	0.307	0.316	0.936
-3	5.84	5.90	0.289	0.295	0.954	7.00	7.08	0.307	0.316	0.936
-2	6.18	6.24	0.297	0.302	0.952	7.27	7.34	0.306	0.314	0.940
-1	6.63	6.68	0.306	0.310	0.952	7.64	7.69	0.301	0.305	0.936
0	7.18	7.22	0.309	0.311	0.954	8.10	8.10	0.292	0.292	0.938
1	7.72	7.76	0.298	0.301	0.948	8.54	8.52	0.278	0.279	0.934
2	8.20	8.23	0.279	0.281	0.944	8.92	8.89	0.259	0.261	0.938
3	8.59	8.60	0.261	0.261	0.946	9.21	9.16	0.240	0.244	0.938
4	8.86	8.84	0.248	0.248	0.948	9.40	9.35	0.225	0.230	0.942
5	9.03	9.00	0.238	0.240	0.942	9.50	9.47	0.215	0.218	0.946