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

Daniel Scharfstein (dscharf@jhsph.edu)
Aidan McDermott (amcdermo@jhsph.edu)
Department of Biostatistics
Johns Hopkins Bloomberg School of Public
615 North Wolfe Street
Baltimore, MD 21205

William Olson (wolson@its.jnj.com)
Janssen Scientific Affairs, LLC
1125 Trenton-Harbourton Road, Titusville, New Jersey 08560-2000

Frank Wiegand (fwiegand@its.jnj.com)
Janssen Pharmaceutical Services ,LLC
700 Route 202 S, Raritan, NJ 08869

December 15, 2013

Abstract

We present a global sensitivity analysis methodology for drawing inference about the mean at the final scheduled visit in a repeated measures study with informative drop-out. We review and critique the sensitivity frameworks developed by Rotnitzky *et al.* (1998, 2001) and Daniels and Hogan (2008). We identify strengths and weaknesses of these approaches and propose an alternative. We illustrate our approach via a comprehensive analysis of the RIS-INT-3 trial.

KEYWORDS: Curse of Dimensionality; Explainable drop-out; Exponential Tilting; G-computation; Identification; Missing at Random; Pattern-Mixture Model; Selection Model

1. INTRODUCTION

In 2010, the National Research Council (NRC) issued a reported entitled ”The Prevention and Treatment of Missing Data in Clinical Trials.” This report, commissioned by the United States Food and Drug Administration, provides 18 recommendations targeted at (1) trial design and conduct, (2) analysis and (3) directions for future research. As inference in the presence of missing data ultimately requires untestable assumptions, Recommendation 15 of the NRC report states

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

Broadly speaking, there are three main types of sensitivity analysis: ad-hoc, local and global. Ad-hoc sensitivity analysis involves analyzing the data using a few different methods (e.g., last observation carried forward, complete-case analysis, mixed models, multiple imputation) and evaluating whether the inferences are consistent. Local sensitivity analysis evaluates how inferences vary in a small neighborhood of a benchmark identification assumption, such as missing at random. It is usually carried out through the computation of partial derivatives. In contrast, global sensitivity analysis considers how inferences vary over a much larger neighborhood of identification assumptions. Chapter 5 of the NRC report emphasizes the global approach.

1.1 Ad-Hoc Sensitivity Analysis

The problem with ad-hoc sensitivity analysis is that the assumptions that underlie these methods can be very strong and for many of these methods unreasonable. More importantly, just because the inferences are consistent does not mean that there are no other reasonable assumptions under which the inference about the treatment effect is different.

1.2 Local Sensitivity Analysis

Ma, Troxel and Heitjan (2005), building on the work of Troxel, Ma and Heitjan (2004), developed an index of local sensitivity to non-ignorable drop-out in longitudinal studies. In their approach, they specify fully parametric models for the outcome process and the drop-out mechanism. In a hazard model for drop-out, the risk of last being seen at a specific visit depends on the (potentially unobserved) outcome scheduled to be collected at the next visit and a perturbation parameter ω

($\omega = 0$ implies missing at random). This is the so-called Diggle and Kenward (1994) model. The index of local sensitivity to non-ignorability is the value of the derivative of the estimator for the parameter of interest from the outcome model with respect to ω, evaluated at $\omega = 0$.

Verbeke *et al.* (2001) use the same models as Ma, Troxel and Heitjan (2005), with the exception that the perturbation parameter is specific to subject i, i.e., ω_i. Using the work of Cook (1986), they develop a method of evaluating the local influence of subject i on inference about the parameters of interest. Specifically, the local influence of subject i is defined as the local change in the profile likelihood displacement curve if drop-out for subject i is not missing at random and drop-out for all other subjects is missing at random.

Copas and Eguchi (2001) create a small expanded neighborhood model around the missing at random model. The neighborhood is created using the Kullback-Leibler divergence metric. They then determine a first-order approximation to the largest asymptotic bias that results from analyzing the data under missing at random as opposed to the expanded model.

1.3 Global Sensitivity Analysis

Rotnitzky, Robins and Scharfstein (1998), Robins, Rotnitzky and Scharfstein (2000), Rotnitzky *et al.* (2001), and Scharfstein, Rotnitzky and Robins (1999) developed semi-parametric global sensitivity analysis strategies for longitudinal studies with monotone missing data. In these papers, hereafter referred to as RRS, the hazard of drop-out at a given time is modeled as an estimable function of the observable history of outcomes and auxiliary covariates through that time as well as a non-identifiable, selection bias function of the outcome scheduled to be measured at the end of the study. For each specified selection bias function, they proposed a class of unbiased estimating functions for estimating the mean outcome at the final visit. Sensitivity analysis is conducted by parameterizing the selection bias function and varying these parameters over ranges considered plausible by subject matter experts. The approach is semi-parametric because no distributional assumptions are placed on the outcomes or auxiliary covariates.

Daniels and Hogan (2008), hereafter DH, specify fully parametric models for the repeated measures within each drop-out pattern. They treat parameters of these models that are not identified as sensitivity analysis parameters. They use fully Bayesian methods to draw inference about the mean outcome at the final visit.

1.4 Our Focus

In this paper, we focus on global sensitivity analysis. In our view, this approach is substantially more informative than the local approach because it (1) allows exploration of the impact of plausible assumptions outside the neighborhood of the benchmark assumption and (2) operates like ”stress testing” in reliability engineering, where a product is systematically subjected to increasingly exaggerated forces/conditions in order to determine its breaking point. Global sensitivity analysis allows one to see how far one needs to deviate from the benchmark assumption in order for inferences to change. If the assumptions under which the inferences change are 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. The approach considered in this paper is similar in spirit to ”tipping-point” analysis (see, for example, Yan, Lee and Li, 2009).

1.5 Outline

In Section 2, we introduce the notation and data structure. We ignore auxiliary covariates in this paper. In Section 3, we review and critique the aforementioned global approaches. In Section 4, we introduce our new approach. In Section 5, we provide a full sensitivity analysis of the RIS-INT-3 trial, a randomized, placebo controlled trial comparing four fixed doses of risperidone, an atypical antipsychotic, and one dose of haloperidol in schizophrenic patients. Section 6 is devoted to a discussion, including how our approach can be extended to incorporate auxiliary covariates.

2. DATA STRUCTURE AND NOTATION

Let $k = 0\text{ (baseline)}, 1, \ldots, K$ be the scheduled assessment times. 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 $R_0 = 1$, and $R_k = 0$ implies $R_{k+1} = 0$, for $k = 1, \ldots, K-1$ (i.e., monotone drop-out). Let $C = \max\{k : R_k = 1\}$, so $C = K$ means that the individual completed the study. We use the notational convention $\overline{Y}_k = (Y_0, \ldots, Y_k)$ and $\underline{Y}_k = (Y_{k+1}, \ldots, Y_K)$. The observed data for an individual is $O = (C, \overline{Y}_C)$. We assume that we observe n i.i.d copies of O, and we want to estimate $\mu = E[Y_K]$.

3. REVIEW AND CRITIQUE OF RRS AND DH

3.1 RRS

RRS showed that μ is non-parametrically identified under the following pattern-mixture modeling assumption:

$$f(Y_K|C=k,\overline{Y}_k)=\frac{f(Y_K|C\geq k+1,\overline{Y}_k)\exp(q_k(\overline{Y}_k,Y_K))}{E[\exp\{q_k(\overline{Y}_k,Y_K)\}\,|\,C\geq k+1,\overline{Y}_k]}\text{ for }k=0,\ldots,K-1,$$

where $q_k(\overline{Y}_k,Y_K)$ is a specified function of Y_K and $\overline{Y}_k$, which is non-identifiable and varied, using subject-matter guidance, in a sensitivity analysis. At each time k, this assumption relates, conditional on past history $\overline{Y}_k$, the distribution of Y_K for those who are last seen at assessment k to those who are on study at $k+1$. When q_k is constant in Y_K for all k, then the model indicates that, conditional on past history $\overline{Y}_k$, individuals who drop-out between assessments k and $k+1$ have the same distribution of Y_K as those who are on study at $k+1$. In this case, RRS refer to drop-out as explainable (or missing at random). If q_k is an increasing (decreasing) function of Y_K for some k, then the model indicates that individuals who drop-out between assessments k and $k+1$ tend to have higher (lower) values of Y_K than those who are on study at $k+1$. In this case, RRS refer to drop-out as non-explainable (or non-ignorable). RRS showed that the above modeling assumption can be expressed in a selection model format as follows:

$$\text{logit}\{P[C=k\,|\,C\geq k,\overline{Y}_k,Y_K]\}=h_k(\overline{Y}_k)+q_k(\overline{Y}_k,Y_K)$$

where

$$h_k(\overline{Y}_k)=\text{logit}\{P[C=k\,|\,C\geq k,\overline{Y}_k]\}-\log\{E[\exp\{q_k(\overline{Y}_k,Y_K)\}\,|\,C\geq k+1,\overline{Y}_k]\}$$

In this form, q_k quantifies the influence of Y_K on the risk of dropping out between assessments k and $k+1$, after controlling for the past history $\overline{Y}_k$.

RRS argued that, due to the curse of dimensionality, additional modeling restrictions were required in order to obtain estimates of μ that converge at rates fast enough (i.e., $\sqrt{n}$) to be of practical use. They further argued that the most flexible way to introduce these restrictions was to parametrically impose modeling assumptions on $h_k(\overline{Y}_k)$. Ultimately, the model they propose places semi-parametric restrictions on the distribution of the observed data.

RRS derived the class of all estimating functions, which involve two pieces. The first piece is the typical inverse weighted estimating function, where individuals with complete data are re-weighted by their probability of completing the study. The second piece is called the augmentation term, which brings in information on those who do not complete the study. Both pieces involve user-specified functions that affect the efficiency of the resulting estimator. The most efficient choice of these functions is very complicated to compute and RRS provide some guidance about how to derive reasonably efficient estimators.

Although the RRS approach has great utility and offers the robustness associated with semi-parametric modeling, there are several aspects of their approach that are dissatisfying. First, we have found that subject matter experts who have been exposed to the RRS technology have difficulty quantifying how the distal outcome scheduled to be measured at the end of the study affects the risk of dropping out at intermediate time points. Rather, we found that these experts were more comfortable thinking about how the outcome scheduled to be measured at assessment $k+1$ affects the risk of dropping out between assessments k and $k+1$. Second, the parametric restrictions on $h_k(\overline{Y}_k)$ induce restrictions on the distribution of the observed data that allow one to, in theory, rule out specific choices of q_k. That is, q_k becomes identifiable. To address this issue, RRS recommend that one impose the weakest parametric restrictions on $h_k(\overline{Y}_k)$, so that (1) μ is $\sqrt{n}$-estimable and (2) the power of testing any specific choice of q_k in finite samples is close to zero. Third, their estimation approach for $h_k(\overline{Y}_k)$ and any associated model selection procedure will be computationally intensive, as it must be performed for each choice of the functions q_k.

3.2 DH

DH specify (a) fully parametric models for $f(\overline{Y}_k|C=k)$, whose parameters are identifiable and (b) fully parametric models for $f(\underline{Y}_k|C=k,\overline{Y}_k)$, whose parameters are not identified. They express the non-identified parameters in terms of identified parameters and sensitivity analysis parameters. The parameter μ can then be expressed in terms of the identifiable parameters and the sensitivity analysis parameters.

For example, in the case where $K=2$ and the outcomes are continuous, DH (Chapter 10) model $f(\overline{Y}_k|C=k)$ in pieces as follows: $Y_0|C=k \sim N(\mu^{(k)},\sigma^{(k)})$, $Y_1|Y_0,C=1$ and $Y_1|Y_0,C=2 \sim N(\alpha_0^{(\geq 1)}+\alpha_1^{(\geq 1)}Y_0,\tau_1^{(\geq 1)})$ and $Y_2|Y_0,Y_1,C=2 \sim N(\beta_0^{(2)}+\beta_1^{(2)}Y_0+\beta_2^{(2)}Y_1,\tau_2^{(2)})$; they model $f(\underline{Y}_k|C=k,\overline{Y}_k)$ in pieces as follows: $Y_1|Y_0,C=0 \sim N(\alpha_0^{(0)}+\alpha_1^{(0)}Y_0,\tau_1^{(0)})$, $Y_2|Y_0,Y_1,C=0 \sim N(\beta_0^{(0)}+\beta_1^{(0)}Y_0+\beta_2^{(0)}Y_1,\tau_2^{(0)})$, $Y_2|Y_0,Y_1,C=1 \sim N(\beta_0^{(1)}+\beta_1^{(1)}Y_0+\beta_2^{(1)}Y_1,\tau_2^{(1)})$. Then they link the non-identified to the identified parameters as follows: $\alpha_s^{(0)} = \alpha_s^{(\geq 1)} + \Delta_{\alpha_s}^{(0:1)}$ for $s = 0, 1$, $\beta_s^{(k)} = \beta_s^{(2)} + \Delta_{\beta_s}^{(k:2)}$ for $s = 0, 1, 2$ and $k = 0, 1$, $\tau_1^{(0)} = \exp\left(\Delta_{\tau_1}^{(0:1)}\right) \tau_1^{(\geq 1)}$, and $\tau_2^{(k)} = \exp\left(\Delta_{\tau_2}^{(k:2)}\right) \tau_2^{(2)}$ for $k = 0, 1$, where the Δ's are sensitivity analysis parameters. When the Δ's are all zero, then missing at random holds. Since the number of sensitivity analysis parameters is large, DH set all parameters equal to zero except $\Delta_{\alpha_0}^{(0:1)}$, $\Delta_{\beta_0}^{(0:2)}$ and $\Delta_{\beta_0}^{(1:2)}$.

In contrast to RRS, an advantage of the modeling approach of DH is that it does not induce identifiable restrictions on non-identified sensitivity analysis parameters. It is also a likelihood-based procedure. Standard goodness of fit procedures can be used to check the adequacy of the models for $f(\overline{Y}_k | C = k)$. The main disadvantage of their approach is that they specify fully parametric models for $f(\underline{Y}_k | C = k, \overline{Y}_k)$ which are not required for identification of μ.

4. PROPOSED METHODOLOGY

Our proposed methodology will address the above concerns about the RRS and DH approaches. In so doing, we introduce alternative assumptions/restrictions that are stronger than those of RRS but weaker than DH. Like DH, we will rely on fully parametric models for the distribution of the observed data.

4.1 Assumptions

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

$$f(Y_K | C = k, \overline{Y}_k, Y_{k+1}) = f(Y_K | C \geq k+1, \overline{Y}_k, Y_{k+1}) \tag{1}$$

This assumption states that, for the cohort patients who are on study at assessment k, share the same history of outcomes through that visit and have the same outcome at assessment $k+1$, the distribution of Y_K is the same for those who are last seen at assessment k and those who are on-study at $k+1$.

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

$$f(Y_{k+1} | C = k, \overline{Y}_k) = \frac{f(Y_{k+1} | C \geq k+1, \overline{Y}_k) \exp\{r_k(\overline{Y}_k, Y_{k+1})\}}{E[\exp\{r_k(\overline{Y}_k, Y_{k+1})\} \,|\, C \geq k+1, \overline{Y}_k]} \tag{2}$$

where $r_k(\overline{Y}_k, Y_{k+1})$ is a specified function of $\overline{Y}_k$ and Y_{k+1}.

This assumption relates, conditional on past history $\overline{Y}_k$, the distribution of Y_{k+1} for those who drop-out between assessments k and $k+1$ to those who are on study at $k+1$. When r_k is constant in Y_{k+1} for all k, then the model indicates that, conditional on past 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 r_k is an increasing (decreasing) function of Y_{k+1} for some k, then the model indicates that 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 $k+1$.

It can be shown that the above modeling assumptions can be expressed in a selection model format as follows:

$$\text{logit}\{P[C = k \,|\, C \geq k, \overline{Y}_{k+1}, Y_K]\} = l_k(\overline{Y}_k) + r_k(\overline{Y}_k, Y_{k+1})$$

where

$$l_k(\overline{Y}_k) = \text{logit}\{P[C = k \,|\, C \geq k, \overline{Y}_k]\} - \log\{E[\exp\{r_k(\overline{Y}_k, Y_{k+1})\} \,|\, C \geq k+1, \overline{Y}_k]\} \tag{3}$$

Notice that $l_k(\overline{Y}_k)$ is identifiable since (1) $P[C = k \,|\, C \geq k, \overline{Y}_k]$ and (2) $f(Y_{k+1}|C \geq k+1, \overline{Y}_k)$ are identified from the distribution of the observed data. In this selection model format, r_k 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 $\overline{Y}_k$ and says that Y_K does not additionally influence this risk. When r_k does not depend on Y_{k+1} the drop-out mechanism is explainable (or missing at random). For specified r_k, Assumptions 1 and 2 place no restrictions on the distribution of observed data. Thus, r_k is not empirically verifiable.

Here, it is important to note that, unlike DH, we do not model $f(Y_{K-1}, \ldots, Y_{k+2}|C = k, \overline{Y}_k)$. It is the imposition of Assumption 1, which allows us to avoid such modeling.

4.2 Identifiability

Given r_k, μ is identifiable. In establishing identifiability, we assume that the distribution of the observed data is known and show that μ can be written as a functional of this distribution. This follows immediately by noting (using repeated application of the law of iterated expectations) that μ

$$\mu = E\left[\frac{I(C = K)Y_K}{\prod_{k=0}^{K-1}(1 + \exp(l_k(\overline{Y}_k) + r_k(\overline{Y}_k, Y_{k+1})))^{-1}}\right], \tag{4}$$

where the expectation is of an observed data random variable. Now, the right hand side of (4) can be re-expressed so that

$$
\begin{aligned}
\mu &= \int_{y_0} \cdots \int_{y_K} \frac{y_K}{\prod_{k=0}^{K-1}(1+\exp(l_k(\overline{y}_k)+r_k(\overline{y}_k,y_{k+1})))^{-1}} \\
&\qquad \prod_{k=0}^{K-1}\left\{dF(y_{k+1}|C\geq k+1,\overline{Y}_k=\overline{y}_k)P[C\geq k+1|C\geq k,\overline{Y}_k=\overline{y}_k]\right\} dF(y_0) \\
&= \int_{y_0}\cdots\int_{y_K} y_K \prod_{k=0}^{K-1}\Big\{dF(y_{k+1}|C\geq k+1,\overline{Y}_k=\overline{y}_k)P[C\geq k+1|C\geq k,\overline{Y}_k=\overline{y}_k]+ \\
&\qquad \frac{dF(y_{k+1}|C\geq k+1,\overline{Y}_k=\overline{y}_k)\exp(r_k(\overline{y}_k,y_{k+1}))}{E[r_k(\overline{Y}_k,Y_{k+1})|C\geq k+1,\overline{Y}_k=\overline{y}_k]})P[C=k|C\geq k,\overline{Y}_k=\overline{y}_k]\Big\} dF(y_0),
\end{aligned} \tag{5}
$$

where all distributions within the integrals are identifiable.

4.3 Estimation and Inference

For given r_k, the identification results above suggest two approaches for estimation of μ. In both approaches, we proceed by specifying fully parametric models for $f(Y_{k+1}|C\geq k+1,\overline{Y}_k)$ (parameters η) and $P[C=k|C\geq k,\overline{Y}_k]$ (parameters γ). The parameters of these models can be estimated using maximum likelihood. Denote these parameter estimates by $\widehat{\eta}$ and $\widehat{\gamma}$, respectively.

Inverse Probability Weighted Estimator (IPW): Following (4), μ can be estimated by

$$
E_n\left[\frac{I(C=K)Y_K}{\prod_{k=0}^{K-1}(1+\exp(l_k(\overline{Y}_k;\widehat{\eta},\widehat{\gamma})+r_k(\overline{Y}_k,Y_{k+1})))^{-1}}\right]
$$

where

$$
l_k(\overline{Y}_k;\eta,\gamma)=\operatorname{logit}\{P[C=k|C\geq k,\overline{Y}_k;\gamma]\}-\log\left\{\int r_k(\overline{Y}_k,y_{k+1})dF(y_{k+1}|C\geq k+1,\overline{Y}_k;\eta)\right\}
$$

and $E_n[\cdot]$ is the empirical expectation operator. This estimator can also be normalized, without affecting its large sample distribution, by dividing by

$$
E_n\left[\frac{I(C=K)}{\prod_{k=0}^{K-1}(1+\exp(l_k(\overline{Y}_k;\widehat{\eta},\widehat{\gamma})+r_k(\overline{Y}_k,Y_{k+1})))^{-1}}\right]
$$

Normalization serves to insure that, if Y_K is a bounded random variable, the resulting estimator will respect these bounds. We will refer to the normalized estimator as the IPW estimator.

The IPW estimator is *not* the maximum likelihood (ML) estimator of μ. This is because it utilizes the empirical expectation operator rather than the expectation operator based on our model for the distribution of the observed data. The G-computation estimator discussed next is the ML estimator of μ.

It is important to notice that in our IPW estimator, we estimated $l_k(\overline{Y}_k)$ indirectly via a model for the distribution of the observed data. In contrast, one could have, like RRS, directly specified a parametric model for $l_k(\overline{Y}_k)$. By adopting this latter approach, one can, in theory, rule out specific choices of r_k, which we know, by construction, is not identifiable. To see why, suppose the model for $l_k(\overline{Y}_k)$ is specified to be linear in Y_k and suppose that for given r_k, there is statistical evidence to suggest that the right hand side of (3) is non-linear in Y_k; then we can reject that choice of r_k. To address this issue, one can, as suggested by RRS, increase the flexibility of the model for $l_k(\overline{Y}_k)$ so that there will be no statistical evidence to reject any choice of r_k. In our view, this is a very cumbersome, as it must be carried out for each specification of r_k.

G-computation Estimator: Following (5), estimation of μ proceeds by repeatedly applying, say 10,000 times, the following simulation procedure and averaging the resulting simulated Y_K's.

1. Simulate Y_0 from its empirical distribution. Set $k = 0$
2. Simulate R_{k+1} from $P[C \geq k+1 | C \geq k, \overline{Y}_k; \widehat{\gamma}]$.
3. If $R_{k+1} = 1$,, simulate Y_{k+1} from $f(Y_{k+1} | C \geq k+1, \overline{Y}_k; \widehat{\eta})$.
4. If $R_{k+1} = 0$, simulate Y_{k+1} from the right hand side of (2)
5. Set $k = k+1$. If $k = K$ then stop; otherwise go to Step 2.

This algorithm is an extension of the G-computation approach developed by Robins (1986) to non-ignorable missing data. Robins, Rotnitzky and Scharfstein (2000) presented a G-computation algorithm for non-ignorable missing data under a stronger set of assumptions.

In the above procedure, one needs to be able to draw from the right hand side of (2). In our data analysis below, we used inverse cumulative distribution function (cdf) sampling. Specifically, we generated a $Uniform(0,1)$ random variable and the computed, using numerical integration (adaptive Gauss Kronrod procedure) and bisection search (Brent-Dekker algorithm), the associated inverse of the cdf.

Both the IPW and G-computation estimators can be shown to be asymptotically normal. The standard error of these estimators and associated confidence intervals can be obtained via non-parametric bootstrap. Treatment effects can be estimated by applying the above procedure separately to each treatment arm.

4.4 Parameterization of $r_k(\overline{Y}_k, Y_{k+1})$

It is not possible to explore all possible choices of $r_k(\overline{Y}_k, Y_{k+1})$. Thus, it is recommended that one consider a low-dimensional parameterization that reflects expert's beliefs about the drop-out process. In the next section, we will choose $r_k(\overline{Y}_k, Y_{k+1}) = \alpha r(Y_{k+1})$, where $r(\cdot)$ is a specified function that serves to quantity the experts' belief that there is a non-linear effect (on the logistic scale) of Y_{k+1} on the risk of dropping out between assessments k and $k+1$ (more later). The parameter α is varied in a sensitivity analysis. To understand the impact of various choices of α, we recommend that one estimate the induced mean of Y_K among drop-outs and compare it to the observed mean among completers. Such a comparison can be used to assess the plausibility of specific choices of α.

5. RIS-INT-3

RIS-INT-3 was a randomized, placebo controlled trial comparing the effectiveness of four fixed doses of risperidone and one dose of haloperidol in schizophrenic patients. 521 patients were randomized in equal proportions to receive 8 weeks of therapy with either 2, 6, 10, 16 mg of risperidone, 20mg haloperidol, or placebo (Marder and Meibach, 1994; Chouinard *et al.*, 1993). The primary outcome of interest was the total Positive and Negative Syndrome Scale score (PANSS score or simply PANSS) after 8 weeks (day 56) of treatment. PANSS ranges from 30 to 210, with higher scores representing greater mental illness.

During the first week of treatment, fixed titration was required to reach the maximal dose within each treatment group. Patients were scheduled to have an assessment performed prior to randomization ($k = 0$). The patients were then assessed during 5 subsequent visits after randomization at days 7 ($k = 1$), 14 ($k = 2$), 28 ($k = 3$), 42 ($k = 4$) and 56 ($k = 5$). The goal was to estimate the mean PANSS at week 56 for each of the 6 arms and contrast the means in the active therapy arms to the placebo arm. For purposes of this illustrative analysis, we focus on the 6mg risperidone ($n = 86$) and placebo ($n = 88$) arms.

There was substantial premature discontinuation of assigned therapy during the course of trial and patients were not subsequently followed. Table 1 shows the treatment-specific cumulative probability of premature withdrawal for the five post-baseline assessment times. Only 31% of patients in the placebo arm completed the study; 58% of patients dropped out due to lack of efficacy. In contrast, 62% of patients in the 6mg risperidone completed the study; 14% of patients dropped out due to lack of efficacy. Other reasons for drop-out included adverse experiences, withdrawal of consent and uncooperativeness. In our analysis, we do not distinguish among the causes of drop-out.

[Table 1 about here.]

Figure 1 displays the treatment-specific trajectory of observed mean PANSS, stratified by last assessment time. It is interesting to note that for patients who prematurely withdraw, the mean PANSS at the last visit tends to be higher than at the previous visit. This is especially dramatic in the placebo arm, consistent with lack of efficacy being the primary reason for premature withdrawal.

[Figure 1 about here.]

[Table 2 about here.]

In our analysis, $K = 5$ and Y_k is PANSS at assessment k. All models were fit separately for each treatment group. We fit the following model for drop-out:

$$\text{logit}P[C = k | C \geq k, \overline{Y}_k] = \gamma_{0,k} + \gamma_1 Y_k \tag{6}$$

For the observed PANSS at assessment $k+1$, we fit a truncated (between 30 and 210) normal regression model of the following form:

$$f(y_{k+1} | C \geq k+1, \overline{Y}_k) = \frac{\phi\left(\frac{y_{k+1} - \eta_{0,k+1} - \eta_{1,k+1} Y_k}{\eta_{2,k+1}}\right)}{\Phi\left(\frac{210 - \eta_{0,k+1} - \eta_{1,k+1} Y_k}{\eta_{2,k+1}}\right) - \Phi\left(\frac{30 - \eta_{0,k+1} - \eta_{1,k+1} Y_k}{\eta_{2,k+1}}\right)} \quad 30 \leq y_{k+1} \leq 210 \tag{7}$$

The treatment-specific parameter estimates (and associated 95% bootstrap confidence intervals) from these models are displayed in Table 2. As the table shows, the PANSS among those on-study at assessment k is positively associated with drop-out between assessments k and $k+1$ in the both treatment arms: the association is stronger in the placebo arm as compared to the 6 mg risperidone arm. Among those on-study at assessment $k+1$, PANSS at assessment k is highly positively predictive of PANSS at assessment k in both treatment arms.

For each treatment group, we evaluated the goodness of fit of the drop-out and outcomes models by computing, respectively, two statistics: $S_1 = E_n[\sum_{k=0}^{4} I(C \geq k)(I(C \geq k+1) - P[C \geq k+1|C \geq k, \overline{Y}_k; \widehat{\gamma}])^2]$ and $S_2 = E_n[\sum_{k=0}^{4} I(C \geq k+1)(Y_{k+1} - E[Y_{k+1}|C \geq k+1, \overline{Y}_k; \widehat{\eta}])^2$. For well fitting models, these statistics should be "small." We estimated the "null distribution" of these statistics by parametric bootstrap. Specifically, we simulated 1,000 datasets under our model-based estimate of the distribution of the observed data. For each simulated dataset, we re-estimated the distribution of the observed data using models (6) and (7) and re-computed the statistics S_1 and S_2. We then computed the p-value associated with each test statistic to be the proportion of the simulated datasets that yielded test statistics greater than the one observed. The p-values for the drop-out model are 0.37 and 0.50 for the placebo and risperidone arms, respectively; for the outcome model the respective p-values are 0.34 and 0.38. For each treatment group, we also compared model-based estimates of the conditional probability of last being seen at visit k given on-study at visit k ($P[C = k|C \geq k]$) to the observed conditional proportions as well as compared model-based estimates of the conditional mean and variance of the outcome at visit $k+1$ given on-study at visit $k+1$ ($E[Y_{k+1}|C \geq k+1]$, $Var[Y_{k+1}|C \geq k+1]$) to the observed conditional means and variances. The model-based and empirical estimates agreed quite well. These analyses suggest that models (6) and (7) provide a reasonable characterization of the distribution of the observed data.

In our analysis, we let

$$r_k(\overline{Y}_k, Y_{k+1}) = \alpha r\left(\frac{Y_{k+1} - 30}{180}\right)$$

where α is a sensitivity analysis parameter and $r(\cdot)$ is the cumulative distribution of a $Beta(4,7)$ random variable (see Figure 2). To understand this choice of selection bias function, consider two patients who are on study through assessment k and have the same history of measured factors through that assessment. Suppose that the first and second patients have PANSS at visit $k+1$ of y_{k+1} and y^*_{k+1}, respectively ($y_{k+1} < y^*_{k+1}$). Then, the logarithm of the odds ratio of last being seen at assessment k as opposed to remaining on study for the second versus the first patient is equal to $\alpha\{r(\frac{y^*_{k+1}-30}{180}) - r(\frac{y_{k+1}-30}{180})\}$. Table 3 below shows the logarithm of the odds ratio for choices of y_{k+1} and y^*_{k+1} that differ by 20 points.

When comparing patients on the very low end or high end of the PANSS scale there is relatively less difference in the risk of drop-out than when comparing patients in the middle of the PANSS scale. Clinical experts refer to this as the floor and ceiling effect of clinical measurement scales. At the extremes of the assessment scale patients are either very well, or extremely sick. The nominal difference in the total value of the disease items which spread across the multiple evaluation domains does not account for a large measurable and clinical meaningful difference in symptomatology on both ends of the scale spectrum. In other words, a very healthy patient does not improve in a clinically meaningfully way if the total PANSS value improves by 20 points and a patient with severe psychosis does not worsen much in the eyes of a clinician if the score increases further. Clinical assessment scales have the tendency to loose discriminatory power at the extremes. These psychometric properties of the scale explain the lack of influence on the odds ratio for discontinuation.

[Figure 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

In what follows, we utilized the G-computation estimation approach. Table 4 displays the treatment-specific mean PANSS at the fifth assessment under the assumption of explainable drop-out ($\alpha = 0$ in each treatment arm). For comparison, we also display the mean PANSS at the fifth assessment for completers. In the placebo group, notice the large difference between the observed mean and the estimated mean under explainable drop-out. In contrast, the difference in the 6mg risperidone arm is much smaller. This is because (1) there is more drop-out in the placebo arm and (2) there is relatively more drop-out due to lack of efficacy in the placebo arm. Under explainable drop-out, there is greater than 18 point difference in the PANSS at the fifth assessment in favor of 6mg risperidone. The result is statistically significant at the 0.05 level as reflected by the fact that the 95% confidence interval for the difference does not contain zero.

For each treatment group, we ranged α from -10 to 25. In the first row of Figure 5, we display the treatment-specific mean PANSS at the fifth assessment as a function of α, along with 95% pointwise confidence intervals. In the bottom row of Figure 5, we display the treatment-specific difference between the mean PANSS at the fifth assessment and the mean PANSS at the fifth assessment among completers, as a function of α, along with 95% pointwise confidence intervals. By viewing these latter figures, subject matter experts can judge the plausibility of various choices of α.

[Figure 3 about here.]

Figure 6 displays a contour plot of the estimated differences between mean PANSS at the fifth assessment for placebo vs. 6mg risperidone for various treatment-specific combinations of α. The dots indicate whether the treatment difference would be statistically significant at the 0.05 level. Over the majority of the plot, the results are statistically (and clinically) significant in favor of 6mg risperidone. Only if the treatment-specific values of the treatment-specific sensitivity analysis parameters are highly differential will the results not be statistically significant. For example, α in the placebo arm would have to be zero and α in the 6 mg risperidone arm would have to be 20. At these values, the difference between the mean PANSS at the fifth assessment and the mean at the fifth assessment among completers would be about 18 and 40 in the placebo and 6 mg risperidone arms, respectively. These differences are not very reasonable, especially given that there is more dropout due to lack of efficacy in the placebo arm. As a result, this sensitivity analysis indicates that inference is robust to deviations from explainable drop-out. That is, 6mg risperidone is superior to placebo in reducing the mean PANSS at the fifth assessment.

[Figure 4 about here.]

We also implemented the IPW estimation approach. In comparison to the G-computation estimation approach, (1) the confidence intervals were noticeably wider and (2) the estimated means as a function of α were non-monotone in the placebo arm. The IPW estimator has the advantage of not extrapolating outside the range of the observed data. In contrast, the G-computation estimation approach does allow for extrapolation. We noticed that when α was largely positive in the placebo arm, the distribution of simulated values of PANSS at the final visit had substantial probability mass placed at values higher than the observed maximum at that visit. This is due, in part, to the small number of patients who complete the study in the placebo arm.

6. DISCUSSION

Our proposed methodology addresses concerns about the RRS sensitivity analysis methodology through imposition of stronger modeling assumptions, both testable and untestable. Our proposal is similar in spirit to that of DH, except that we require weaker modeling assumptions. The advantages of our proposed procedures are four-fold:

1. r_k remains non-identifiable

2. the parameters that govern the model for the observed data can be estimated, via maximum likelihood, independent of r_k

3. the models for the observed data can be developed by an independent data analyst

4. inference for μ is fully efficient, under correct model specification

The disadvantage of the proposed procedure is that it is fully parametric and can therefore be sensitive to model misspecification. This concern is mitigated by the fact the models can be built using relatively standard goodness of fit procedures.

The proposed methodology can be extended to incorporate time-dependent auxiliary variables. The importance of including these auxiliaries is to make the benchmark explainable dropout assumption more tenable. If we let V_k denote the auxiliary variables scheduled to be collected at assessment k and $W_k = (Y_k, V_k)$, then $\overline{Y}_k$ on the right hand side of Equations (1)-(3) would be replaced by $\overline{W}_k = (W_0, \ldots, W_k)$. In these new expressions, when r_k does not depend on Y_{k+1}, one is assuming that for the cohort patients who are on study at assessment k and share the same history of outcomes *and* auxiliaries through that visit, the risk of dropping out before next assessment does not depend on either Y_{k+1} or Y_K. In adapting our likelihood-based approach to incorporate auxiliaries, the distribution of the observed data would require more modeling. Specifically, we would need to model $f(V_{k+1}|C \geq k+1, Y_{k+1}, \overline{W}_k)$. In future work, we will focus on relaxing distributional assumptions on the law of the observed data, while preserving the advantages described above.

The first two authors have been funded by the Food and Drug Administration as well as the Patient-Centered Outcomes Research Institute to develop software for conducting global sensitivity analysis of randomized trials with missing outcome data. The software will be available at `www.missingdatamatters.org`.

REFERENCES

Chouinard, G., Jones, B., Remington, G., Bloom, D., Addinfton, D., MacEwan, G., Labelle, A., Beauclair, L., & Arnott, W. (1993), "A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients," *Journal of Clinical Psychopharmacology*, 13, 25–40.

Cook, D. (1986), "Assessment of Local Influence," *Journal of the Royal Statistical Society, Series B*, 2, 133–169.

Copas, J., & Eguchi, S. (2001), "Local Sensitivity Approximations for Selectivity Bias," *Journal of the Royal Statistical Society, Series B*, 63(871-895).

Daniels, M., & Hogan, J. (2008), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis* CRC Press.

Diggle, P., & Kenward, M. (1994), "Informative Drop-Out in Longitudinal Data Analysis," *Applied Statistics*, 43, 49–93.

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

Ma, G., Toxel, A., & Heitjan, D. (2005), "An Index of Local Sensitivity to Nonignorable Drop-out in Longitudinal Modelling," *Statistics in Medicine*, 24, 2129–2150.

Marder, S., & Meibach, R. (1994), "Risperidone in the treatment of schizophrenia," *American Journal of Psychiatry*, 151, 825–835.

Robins, J. (1986), "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 7, 1393–1512.

Robins, J., Rotnitzky, A., & Scharfstein, D. (2000), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models for Epidemiology*, ed. E. Halloran Springer-Verlag, pp. 1–94.

Rotnitzky, A., Robins, J., & Scharfstein, D. (1998), "Semiparametric Regression for Repeated Outcomes with Non-Ignorable Non-Response," *Journal of the American Statistical Association*, 93, 1321–1339.

Rotnitzky, A., Scharfstein, D., Su, T., & Robins, J. (2001), "A Sensitivity Analysis Methodology for Randomized Trials with Potentially Non-Ignorable Cause-Specific Censoring," *Biometrics*, 57, 103–113.

Scharfstein, D., Rotnitzky, A., & Robins, J. (1999), "Adjusting for Non-ignorable Drop-out Using Semiparametric Non-response Models (with discussion)," *Journal of the American Statistical Association*, 94, 1096–1146.

Troxel, A., Ma, G., & Heitjan, D. (2004), "An Index of Local Sensitivity to Nonignorability," *Statistica Sinica*, 14, 1221–1237.

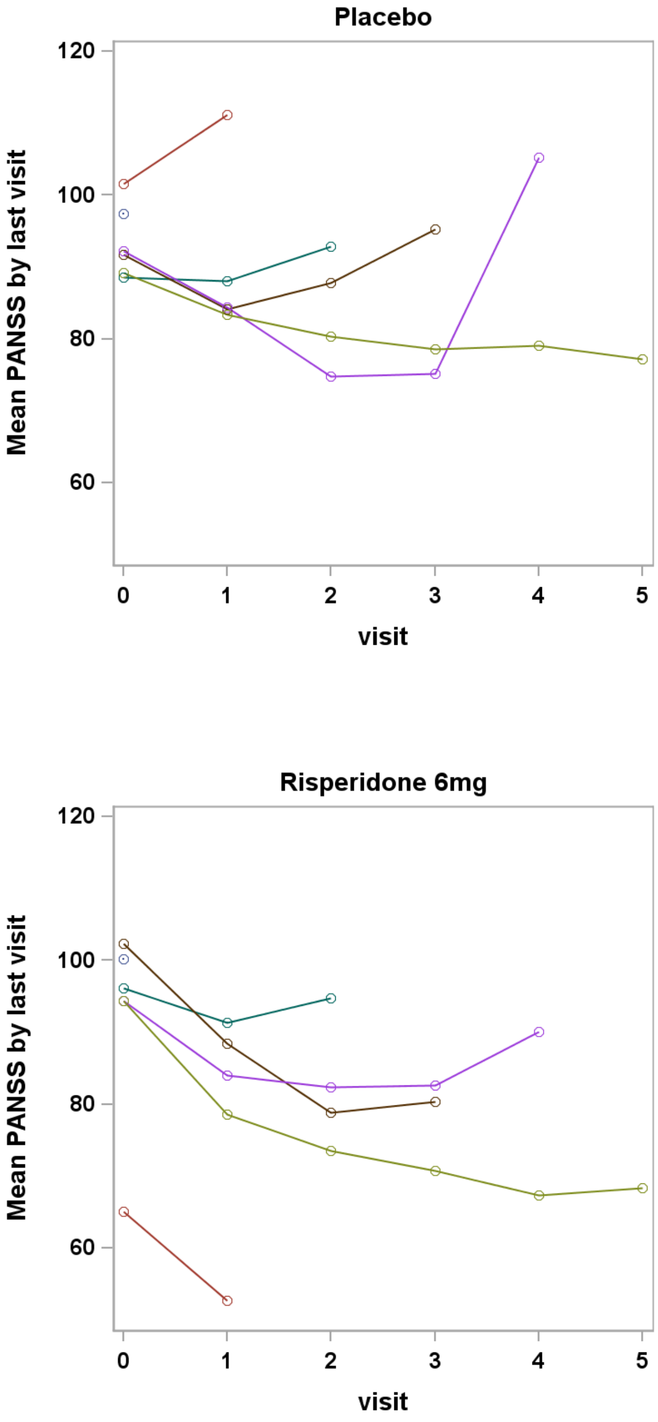
Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., & Kenward, M. (2001), "Sensitivity Analysis for Nonrandom Dropout: A Local Influence Approach," *Biometrics*, 57, 7–14.

Yan, X., Lee, S., & Li, N. (2009), "Missing Data Handling Methods in Medical Device Clinical Trials," *Journal of Biopharmaceutical Statistics*, 19, 1085–1098.

List of Figures

1 Treatment-specific trajectory of observed mean PANSS, stratified by last assessment time . . . 21
2 Kernel of selection bias function . . . 22
3 First row: Treatment-specific mean PANSS at the fifth assessment as a function of α, along with 95% pointwise confidence intervals. Second Row: Treatment-specific difference between the mean PANSS at the fifth assessment and the mean PANSS at the fifth assessment among completers, as a function of α, along with 95% pointwise confidence intervals. . . . 23
4 Contour plot of the estimated differences between mean PANSS at the fifth assessment for placebo vs. 6mg risperidone for various treatment-specific combinations of α. Dots indicate whether the treatment difference would be statistically significant at the 0.05 level. . . . 24

Figure 1: Treatment-specific trajectory of observed mean PANSS, stratified by last assessment time



Placebo

Risperidone 6mg

Figure 2: Kernel of selection bias function

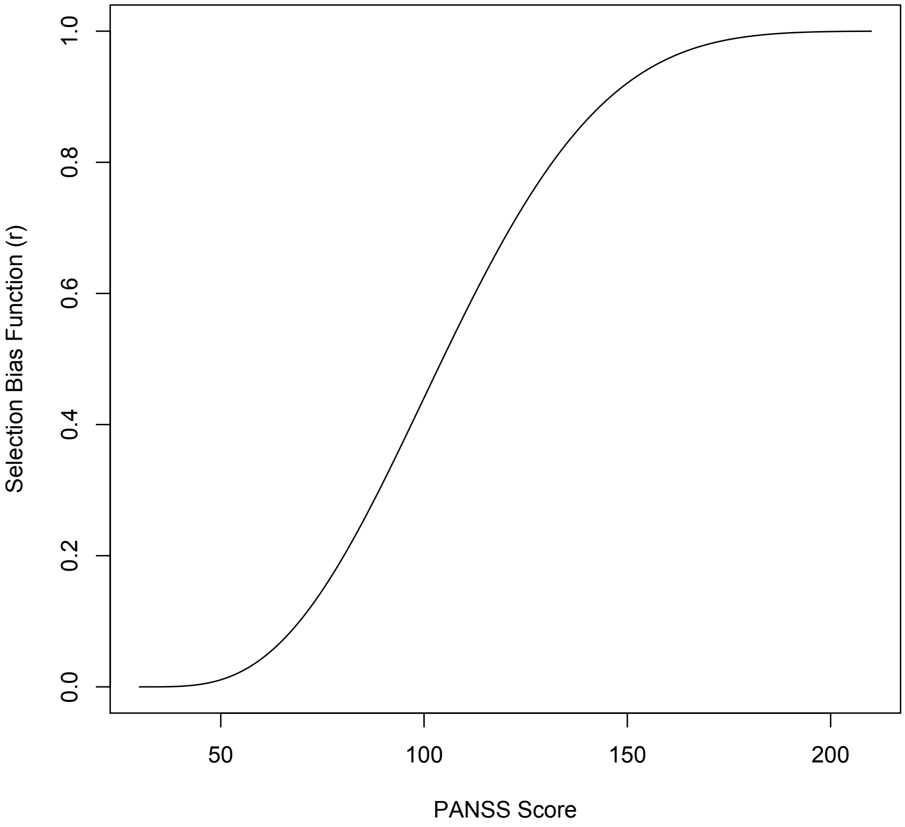


Figure 3: First row: Treatment-specific mean PANSS at the fifth assessment as a function of α, along with 95% pointwise confidence intervals. Second Row: Treatment-specific difference between the mean PANSS at the fifth assessment and the mean PANSS at the fifth assessment among completers, as a function of α, along with 95% pointwise confidence intervals.

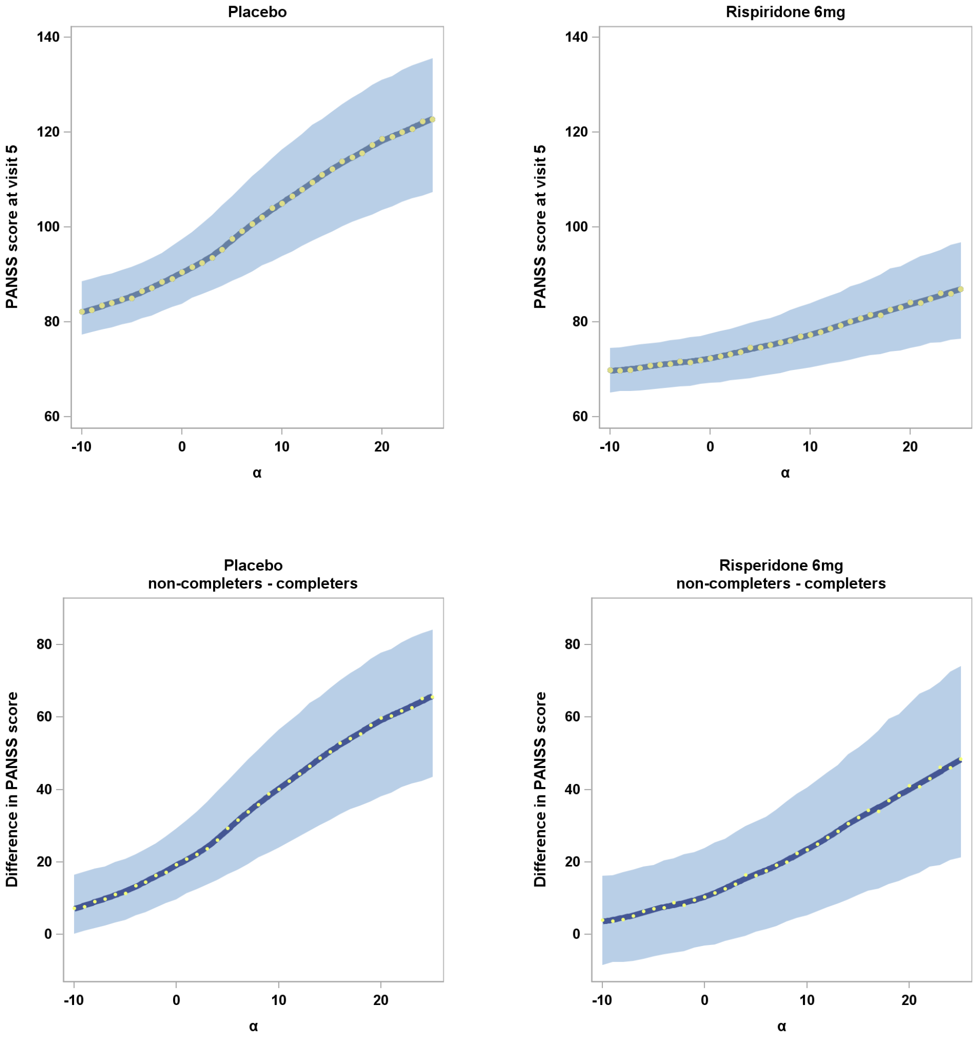
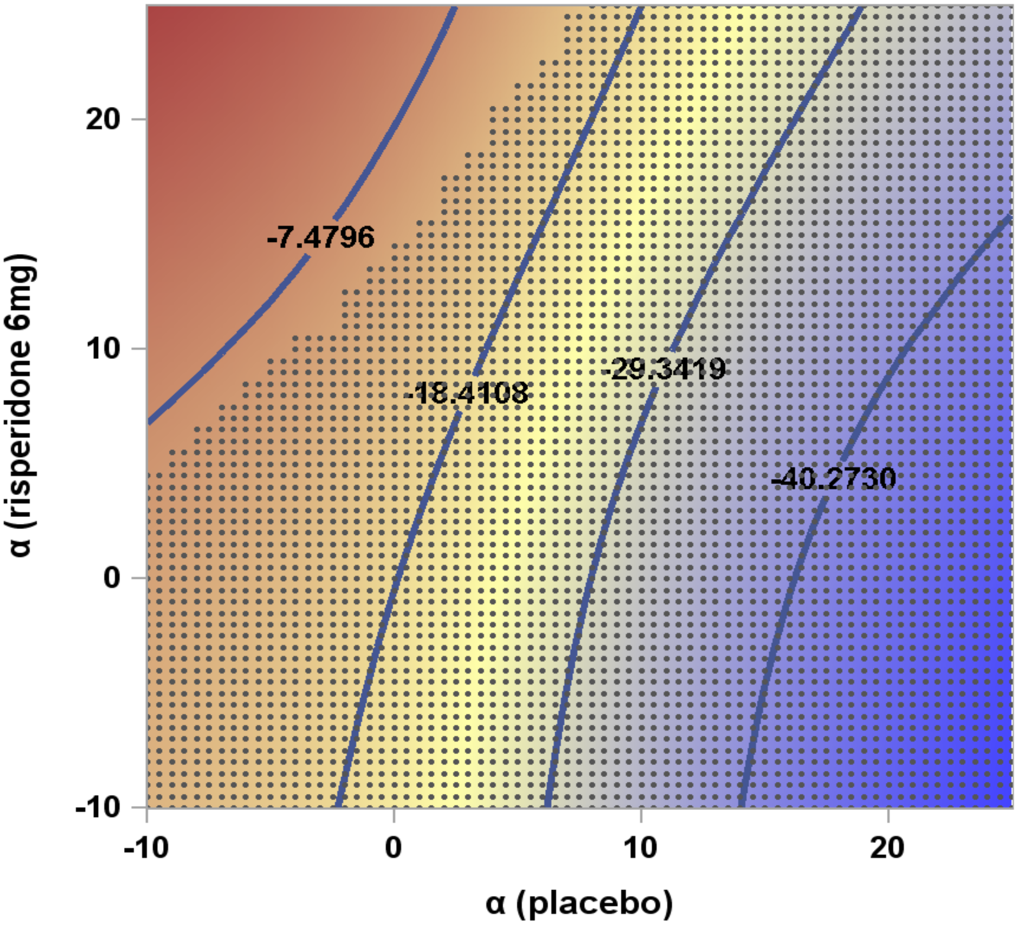


Figure 4: Contour plot of the estimated differences between mean PANSS at the fifth assessment for placebo vs. 6mg risperidone for various treatment-specific combinations of α. Dots indicate whether the treatment difference would be statistically significant at the 0.05 level.



List of Tables

1 Treatment-specific cumulative probability of treatment termination 26

2 Parameter estimates (and 95% confidence intervals) for observed data models: drop-out and PANSS 27

3 The logarithm of the odds ratio of last being seen at assessment k (among those on-study at assessment k and who share the same history of observed data through that assessment) for choices of y_{k+1} and y^*_{k+1} that differ by 20 points. 28

4 Treatment-specific mean PANSS at the fifth assessment for completers and under the assumption of explainable drop-out ($\alpha = 0$ in each treatment arm). 29

Table 1: Treatment-specific cumulative probability of treatment termination

		Cum. prob. of treatment termination				
Treatment	n	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
PLA	88	0.11	0.27	0.49	0.62	0.69
RIS 6mg	86	0.06	0.09	0.22	0.35	0.38

Table 2: Parameter estimates (and 95% confidence intervals) for observed data models: drop-out and PANSS

Dropout model:

	Placebo			**Risperidone 6mg**		
Variable	**Estimate**	**95% CI**		**Estimate**	**95% CI**	
Visit 1 ($\gamma_{0,1}$)	-6.34	-8.94	-4.62	-5.14	-8.39	-2.93
Visit 2 ($\gamma_{0,2}$)	-5.72	-7.97	-4.22	-5.28	-17.88	-2.89
Visit 3 ($\gamma_{0,3}$)	-4.73	-6.80	-3.29	-3.73	-5.63	-2.15
Visit 4 ($\gamma_{0,4}$)	-4.82	-6.77	-3.44	-3.44	-5.42	-1.84
Visit 5 ($\gamma_{0,5}$)	-5.48	-7.62	-4.13	-4.62	-17.01	-2.98
PANSS (γ_1)	0.044	0.029	0.066	0.024	0.003	0.045

PANSS model:

		Placebo			**Risperidone 6mg**		
Outcome	**Variable**	**Estimate**	**95 % CI**		**Estimate**	**95 % CI**	
PANSS$_{t=1}$	Intercept ($\eta_{0,1}$)	11.45	-10.68	30.43	21.47	4.38	39.83
	PANSS$_{t=0}$ ($\eta_{1,1}$)	0.85	0.64	1.08	0.63	0.43	0.82
	Std. Dev. ($\eta_{2,1}$)	15.25	12.46	17.38	14.96	12.09	17.02
PANSS$_{t=2}$	Intercept ($\eta_{0,2}$)	16.80	-0.39	32.80	6.32	-4.17	17.92
	PANSS$_{t=1}$ ($\eta_{1,2}$)	0.80	0.62	1.01	0.87	0.73	1.01
	Std. Dev. ($\eta_{2,2}$)	13.24	10.10	15.65	11.64	9.60	13.45
PANSS$_{t=3}$	Intercept ($\eta_{0,3}$)	14.33	-4.46	33.56	7.68	-5.99	20.20
	PANSS$_{t=2}$ ($\eta_{1,3}$)	0.84	0.61	1.07	0.87	0.72	1.05
	Std. Dev. ($\eta_{2,3}$)	13.04	10.00	15.36	13.48	10.42	16.17
PANSS$_{t=4}$	Intercept ($\eta_{0,4}$)	23.57	2.75	53.53	-4.11	-17.99	10.10
	PANSS$_{t=3}$ ($\eta_{1,4}$)	0.77	0.44	1.00	1.00	0.79	1.20
	Std. Dev. ($\eta_{2,4}$)	17.59	8.66	26.28	12.27	9.20	14.78
PANSS$_{t=5}$	Intercept ($\eta_{0,5}$)	-2.73	-12.75	7.32	5.67	-0.88	14.07
	PANSS$_{t=4}$ ($\eta_{1,5}$)	1.01	0.89	1.16	0.93	0.81	1.02
	Std. Dev. ($\eta_{2,5}$)	7.27	3.55	9.68	6.82	4.64	8.71

Table 3: The logarithm of the odds ratio of last being seen at assessment k (among those on-study at assessment k and who share the same history of observed data through that assessment) for choices of y_{k+1} and y^*_{k+1} that differ by 20 points.

y^*_{k+1}	y_{k+1}	Log Odds Ratio
50	30	$\alpha 0.02$
60	40	$\alpha 0.07$
80	60	$\alpha 0.22$
100	80	$\alpha 0.30$
120	100	$\alpha 0.24$
140	120	$\alpha 0.12$
160	140	$\alpha 0.04$
180	160	$\alpha 0.01$
200	180	$\alpha 0.00$

Table 4: Treatment-specific mean PANSS at the fifth assessment for completers and under the assumption of explainable drop-out ($\alpha = 0$ in each treatment arm).

	Observed	Explainable Drop-out	
	Mean	Estimate	95% CI
Placebo	77.19	90.52	[83.82,97.43]
6mg Risperidone	68.36	72.30	[67.13,77.47]
Difference		-18.22	[-26.50,-9.22]