



Published in final edited form as:

Biometrics. 2010 June ; 66(2): 558–566. doi:10.1111/j.1541-0420.2009.01290.x.

A global sensitivity test for evaluating statistical hypotheses with non-identifiable models

D. Todem^{1,*}, J. Fine², and L. Peng³

¹Division of Biostatistics, Department of Epidemiology, Michigan State University, B601 West Fee Hall, East Lansing, MI 48823, U.S.A.

²Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, 600 Highland Ave., Madison, WI 53792, U.S.A.

³Department of Biostatistics, Rollins School of Public Health Emory University, 1518 Clifton Rd NE

SUMMARY

We consider the problem of evaluating a statistical hypothesis when some model characteristics are non-identifiable from observed data. Such scenario is common in meta-analysis for assessing publication bias and in longitudinal studies for evaluating a covariate effect when dropouts are likely to be non-ignorable. One possible approach to this problem is to fix a minimal set of sensitivity parameters conditional upon which hypothesized parameters are identifiable. Here, we extend this idea and show how to evaluate the hypothesis of interest using an infimum statistic over the whole support of the sensitivity parameter. We characterize the limiting distribution of the statistic as a process in the sensitivity parameter, which involves a careful theoretical analysis of its behavior under model misspecification. In practice, we suggest a nonparametric bootstrap procedure to implement this infimum test as well as to construct confidence bands for simultaneous pointwise tests across all values of the sensitivity parameter, adjusting for multiple testing. The methodology's practical utility is illustrated in an analysis of a longitudinal psychiatric study.

Keywords

Data missing not at random; Functional estimators; Infimum statistic; Model misspecification; Nonparametric bootstrap; Profile estimators; Sensitivity analysis; Uniform convergence

1. Introduction

Sensitivity analysis has been widely advocated in missing data scenarios where the missingness process depends on the unobserved outcomes, as defined in Little and Rubin (1987). Under such missingness mechanism, missing data cannot simply be ignored and unbiased inferences may require strong modelling assumptions, which cannot be checked using the observed data. We consider parametric models, where there are certain parameters of interest in the model for the data and other parameters in the missingness model which may be viewed as secondary parameters. In addition to the form of the parametric model not

*todem@msu.edu.

SUPPLEMENTARY MATERIALS

Web Appendices referenced in Sections 2 and 3 are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

being identifiable, in many cases, the models may be "overparameterized", so that even if the form of the model is correct, standard estimation techniques may still fail due to the models being non-identifiable. This is problematic, since it may not be known a priori whether identifiability holds, owing to the complex nonlinear structure of the missing data models. In practice, these problems may be manifested in unstable estimation, where the algorithms used to compute the estimators may not converge reliably.

In order to make inference about the parameters of interest without simultaneously estimating all model parameters, one may instead fix a minimal set of parameters, called sensitivity parameters, conditional upon which the primary parameters are assumed identifiable. These profile models, viewed as functionals of the sensitivity parameters, provide the basis of sensitivity analysis. In essence, one uses the model profiled across the sensitivity parameters as a working model to delineate inferences consistent with the observed data. Two general strategies for sensitivity analysis have been discussed in the literature, local and global, depending on the range of the sensitivity parameter under consideration. A local sensitivity approach assesses the impacts of uncertainties on inferences over a range of models specified by the sensitivity parameter in a small "neighborhood" of a known value of this parameter (Copas and Eguchi, 2001; Verbeke et al., 2001; Troxel et al., 2002; and Todem et al., 2006). The idea stems from work on missing data, where one may assess the effects of small perturbations of the ignorable model in the directions of non-ignorable models. The sensitivity of the identifiable parameter with respect to sensitivity parameters may be estimated via partial derivatives with respect to the sensitivity parameter in a neighborhood of the sensitivity parameter values corresponding to ignorable missingness. If the parameter estimates are locally insensitive to these values of sensitivity parameters, then, under the assumption of small violations of ignorability, the hypothesis can be reasonably evaluated assuming ignorable missingness. Such methodology is useful but does not permit assessments of large deviations of sensitivity parameters on inferences. In practice, it may be of scientific interest to understand the impact of such deviations. Global analyses may be critically important in such applications. Under qualitative assumptions regarding the missingness mechanism, it may be possible to derive such bounds without imposing further parametric assumptions; see Manski (2003) and Horowitz and Manski (2006) for overviews. Such ideas have been explored empirically, as in Scharfstein, Manski and Anthony (2004) for observational studies. Alternatively, one may employ more structured parametric models for the missingness mechanism. Such modelling approaches are commonly employed in practical data analytic settings, owing to their conceptual simplicity and ease of implementation. Similarly to our set-up, there are a small number of sensitivity parameters. Recent illustrations of such methodology can be found in Shepherd et al (2006) and Vansteelandt et al (2006). A limitation is that formal inference is not provided. The usual strategy for presenting results is to provide parameter estimates at various values of the sensitivity parameters, accompanied by pointwise confidence intervals based on the assumption of a correctly specified model. Inference regarding significance of the covariates is ad hoc, ignoring that multiple tests are conducted and that the model may be misspecified.

In Section 2, we propose conservatively testing the null hypothesis across the support of the sensitivity parameter using an infimum test statistic. The test accounts for the facts that inferences are carried out simultaneously across the entire range of the sensitivity parameter and that the model is misspecified under certain values of the sensitivity parameter, as there is only a single true value. We also develop simultaneous confidence bands for the identifiable parameter, enabling pointwise tests which control the overall type I error rate and an assessment of the magnitudes of the effects, over the range of the sensitivity parameter. These procedures require a careful theoretical analysis of the estimator as a process in the sensitivity parameter; asymptotic properties are sketched in the text, with

details in a supplementary web file. A bootstrap procedure is suggested for practical implementation. In Section 3, we present a special case of our general working model set-up, a shared random effects model for longitudinal binary outcomes with non-random dropouts. We discuss identifiability issues related to the proposed model and describe aspects of estimation and inference. A practical illustration of the sensitivity analysis for this model using a psychiatric data set is given. We conduct a simulation study assessing the performance of the sensitivity tests in small samples. A discussion concludes in section 4. Additional technical details are contained in a Supplementary Materials file.

2. The general framework

2.1 Overview

Consider a random process of any finite dimension $W \sim f_W(\theta; W)$, where $f_W(\theta; \cdot)$ is a parametric family known up to a parameter vector $\theta \in \Theta$. As an example, in the context of meta-analysis on publication bias one may take $W = (Y, R)$, where Y is observed for $R = 0$ and missing for $R = 1$. In the context of longitudinal studies with T potential outcomes for each subject, one would take $Y = (Y_1, \dots, Y_T)'$, and $R = (R_1, \dots, R_T)'$, where $R_t = 1$ if Y_t is unobserved and 0 if otherwise. Suppose that n independent copies W_1, W_2, \dots, W_n of W are available and that consistent estimation of θ may not be possible as a result of identifiability issues. That is, the maximum likelihood estimate of θ namely,

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \ln f_w(\theta, W_i),$$

may not have good properties. The nonidentifiability of the model can be managed by considering a parameter say δ , conditional upon which the others are estimable. The model parameter vector can then be partitioned as $\theta = (\psi, \delta)$, where ψ is assumed identifiable for a fixed δ . The usual strategy for presenting results is to provide the maximum likelihood parameter estimates $\hat{\psi}(\delta)$ of ψ at various values of δ , accompanied by pointwise confidence intervals based on the assumption of a correctly specified model. Inference regarding significance of identifiable model parameters is ad hoc, ignoring that multiple tests are conducted and that the model may be misspecified. Previous work on complex parametric models has overlooked these two key issues, the first being that inference must be made simultaneously across a continuum of the sensitivity parameter and the second being that the missingness model will be misspecified at virtually all values of the sensitivity parameter within that continuum. The theoretical analysis below provides the necessary results for dealing with these issues. In particular, the development of the sensitivity tests in the next subsection depends critically on a careful study of the properties of the $\hat{\psi}(\delta)$.

2.2 Uniform asymptotic properties of profile likelihood-based estimators

To ensure that the profile likelihood function is bounded for any value of δ , we restrict \mathcal{H} , the support of δ , to a compact set. It is well known that under a correctly specified model fit with $\delta = \delta_0$, the true value of δ , and under mild regularity conditions, the estimators are consistent and asymptotically normal, that is,

$$n^{\frac{1}{2}} \{ \hat{\psi}(\delta_0) - \psi_0 \} \rightarrow_d N(0, \Sigma) \text{ and } \hat{\psi}(\delta_0) \rightarrow_p \psi_0$$

as the sample size n gets large, where ψ_0 is the true value of ψ with δ fixed at δ_0 . Under a misspecified dropout process, i.e. $\delta \neq \delta_0$, and for a large n , we get $\hat{\psi}(\delta) \rightarrow_p \psi_*$, with ψ_* not necessarily equal to ψ_0 . The vector ψ_* will typically depend on δ and will be denoted $\psi_*(\delta)$

to highlight this dependence. When $n \rightarrow \infty$, the pseudoscore functions are roughly quadratic in the neighborhood of $\psi_*(\delta)$ for fixed δ and the limiting distribution of $n^{1/2}\{\hat{\psi}(\delta) - \psi_*(\delta)\}$ is mean zero normal with variance-covariance matrix $\Gamma(\delta)$ (White, 1982).

These pointwise results may be made uniform under certain smoothness conditions, using empirical process arguments. Under conditions C1 and C2 in the Supplementary Materials, $\hat{\psi}(\delta)$ is uniformly consistent for $\psi_*(\delta)$ and $J(\delta) = n^{1/2}\{\hat{\psi}(\delta) - \psi_*(\delta)\}$ converges weakly to a Gaussian process with mean 0 and covariance function $E\{J(\delta_1)J(\delta_2)'\} = \Gamma(\delta_1, \delta_2) = E\{\iota_i(\delta_1)\iota_i(\delta_2)'\}$, where the influence function $\iota_i(\delta)$ is defined in (A3) in the Supplementary Materials. In general, estimates at different values of the sensitivity parameter will be highly correlated and treating the estimates as independent could lead to misleading inferences. An estimate of $\Gamma(\delta)$ based on the inverse of the observed information matrix from the profile likelihood is known to be biased, due to the model misspecification. McCullagh and Tibshirani (1990) have shown that some adjustments can be made in order to reduce the bias to $O(1/n)$. An alternative is to use a robust sandwich variance estimator as described by

White (1982), which can be computed as $\hat{\Gamma}(\delta_1, \delta_2) = n^{-1} \sum_i \hat{\iota}_i(\delta_1) \hat{\iota}_i'(\delta_2)$ where $\hat{\iota}_i$ is ι_i with all unknown quantities replaced by empirical estimators. One may also use the bootstrap method to estimate the standard errors of the estimates of the parameter $\psi_*(\delta)$ (Efron and Tibshirani, 1993). The method consists of randomly selecting S samples of size n with replacement from the original data set, where the sampling units consist of all information collected on each individual, $W_i, i = 1, \dots, n$. The sample standard deviation of the parameter estimates from the n bootstrap samples estimates the true standard error. This bootstrap procedure may also be used to approximate the distribution of $\hat{\psi}(\delta)$ as a process, which is quite complex and does not lead to simple analytic testing procedures. The validity of the bootstrap follows automatically from empirical process theory under the regularity conditions given in Appendix A (van der Vaart and Wellner, 2000), essentially requiring the smoothness and boundedness of the likelihood for fixed $\delta \in \mathcal{H}$. This occurs because even though the model may be misspecified, the estimand $\psi_*(\delta)$ is an implicit functional of the empirical distribution so that empirical process theory applies equally under correctly specified and misspecified models.

Condition C2 in the Supplementary Materials underlying all of the above results requires that a unique $\psi_*(\delta)$ exists after fixing δ . In fact, to our knowledge, a rigorous analysis of $\hat{\psi}(\delta)$ accounting for model misspecification has not previously been undertaken. Formally establishing this condition is beyond the scope of this paper.

2.3 Global sensitivity testing

Suppose we are interested in evaluating the following null hypothesis:

$$H_{01}: C\psi = c,$$

where C is an $r \times p$ contrast matrix and c is an $r \times 1$ vector of constants. This framework allows for composite hypotheses. In the special case of testing whether the j th coefficient of ψ is equal to 0, one takes C to be $1 \times p$ vector with a one in the j th position and zeros elsewhere and $c = 0$. Under a nonidentifiable model, the above hypothesis cannot be tested without additional restrictions. If the true sensitivity parameter δ_0 is known, the true hypothesis to be evaluated is given by $H_{02} : C\psi_*(\delta_0) = c$, where $\psi_*(\delta_0) = \psi_0$. Unfortunately in most cases, δ_0 is unknown and may not be estimable from observed data. We propose a global sensitivity test which uses the trivial inequalities given by,

$$\inf_{\delta} \|C\psi_*(\delta) - c\| \leq \|C\psi_*(\delta_0) - c\| \leq \sup_{\delta} \|C\psi_*(\delta) - c\|,$$

to make rigorous inferential statements about H_{02} . Here $\|\cdot\|$ is any norm function, typically Euclidean norm. We can evaluate H_{02} by conservatively testing the hypothesis $\inf_{\delta} \|C\psi_*(\delta) - c\| = 0$. Clearly when $\inf_{\delta} \|C\psi_*(\delta) - c\|$ is strictly greater than zero, $\|C\psi_*(\delta_0) - c\|$ will be greater than zero as well. The infimum hypothesis is formally defined as

$$H_{03} : \inf_{\delta} \|C\psi_*(\delta) - c\| = 0. \tag{1}$$

In the case of a one-dimensional parameter ψ , an ad hoc test may be conducted by constructing simultaneous confidence bands for $\psi_*(\delta)$ given δ . If the band includes 0 for any δ , then H_{02} cannot be rejected. In general, if a simultaneous confidence band for $C\psi_*(\delta)$ excludes c for all δ , then one rejects the associated null. Note that such bands also identify those δ at which the null is rejected, which may be useful in understanding how inferences change as a function of the missingness model.

We now propose a formal infimum statistic to evaluate H_{03} ,

$$T = \inf_{\delta} \left\{ (C\hat{\psi}(\delta) - c)' (C\hat{\Sigma}(\delta)C')^{-1} (C\hat{\psi}(\delta) - c) \right\}$$

where $\hat{\Sigma}(\delta)$ is an estimator of the variance-covariance matrix of $\hat{\psi}(\delta)$. The test statistic T rejects the null for unusual large values. While for each fixed δ , the test process follows a simple chi square distribution under the null, the asymptotic distribution of the estimator $\hat{\psi}(\delta)$ as a process in δ is quite complicated and the distribution of T is analytically intractable. One may use the nonparametric bootstrap to generate the distribution of the infimum test statistic and to construct the confidence bands, using the approach described in section 2.2. Let $\hat{\psi}^s(\delta)$ and T^s , $s = 1, \dots, S$ denote the estimators and infimum tests computed in the S bootstrap samples. One rejects the null at level α if the observed test statistic is larger than the $(1 - \alpha)$ percentile of empirical distribution of T^s , $s = 1, \dots, S$. A simultaneous confidence region for $C\psi_*(\delta) - c$ takes the form

$$\{x(\delta) : \mathcal{H} \rightarrow R^r; \|x(\delta) - C\hat{\psi}(\delta) + c\| < \hat{\vartheta}_{\alpha}\},$$

where $\hat{\vartheta}_{\alpha}$ as the $(1 - \alpha)$ th empirical percentile of $\{\sup_{\delta \in \mathcal{H}} \|C\hat{\psi}^s(\delta) - C\hat{\psi}(\delta)\|\}_{s=1}^S$.

A pointwise approach might also be used to perform a sensitivity analysis for $\|C\psi_*(\delta) - c\| = 0$ given a finite number of values for δ . The method consists of letting the sensitivity parameter take values in a set $A = \{\delta_1, \dots, \delta_Q\} \subset R^+$, with $Q < \infty$ and evaluating the hypothesis, $\|C\psi_*(\delta) - c\| = 0$. Specifically, for $\delta \in A$, simultaneous pointwise confidence intervals may be constructed for $\{C\psi_*(\delta) - c\}$ using standard error estimates for $\hat{\psi}(\delta)$ from section 2.2, with a multiplicity adjustment. If all intervals exclude 0, then $C\psi_*(\delta) \neq c$ at those δ . This approach should be carefully undertaken as the choice of points in A may be somewhat arbitrary and may miss δ where H_{03} holds. Thus, one cannot formally test H_{01} using finite δ . Moreover, there may be reduced power with large Q , where the multiplicity adjustment for controlling the overall type I error may be quite conservative. The global approach described above provides a systematic method for dealing with these issues.

3. Application to longitudinal outcomes with potentially non-random dropouts

3.1 A shared random effects model as a working model

We consider a longitudinal study where for each subject $i = 1, \dots, n$, there are T potential outcomes Y_{i1}, \dots, Y_{iT} represented by the vector $Y_i(\mathcal{C}) = (Y_{i1}, \dots, Y_{iT})'$, measured at discrete time points in $\mathcal{C} = \{t_1, \dots, t_T\}$. These outcomes, however, may not be fully observed and therefore are coupled with a missingness indicator $R_i = (R_{i1}, \dots, R_{iT})'$, where $R_{it} = 1$ if Y_{it} is unobserved and 0 if otherwise. When the missing data result from dropouts, it is intuitive to represent the series of missingness indicators by a single random variable denoted by $D_i = 1 + \max\{t : R_{it} = 0\}$, which indicates the drop-out time for subject i . We assume that $D_i \geq 2$ which implies that all subjects are present at first time point. If $D_i = T + 1$ then the subject is fully observed and if $D_i < T + 1$ the subject drops out.

Although our analysis can be performed for a wide variety of models, to minimize incidental technicalities, we consider the special case of random intercept models with potentially not at random missing values. We consider a conditional mean model, $\mu_{it} = \text{pr}(Y_{it} = 1 \mid b_i, X_{it})$, for an hypothetical response Y_{it} given some random effects b_i as,

$$\mu_{it} = g^{-1}(X_{it}\beta + b_i).$$

Here g is a monotone, differentiable and invertible function, and $\beta = (\beta_1, \dots, \beta_p)$ is the slope vector associated with fixed covariate X_{it} . The random variable b_i is assumed to be generated from a central normal distribution with variance τ^2 . This mean model is suitable for analyzing data for which the primary interest is the assessment of subject's level contrasts. It is well known that this family of models produces fixed effects parameters that have a subject-specific interpretation.

We allow the discrete dropout hazard, $h_{itj} = \text{pr}(t_{D_i} = t_j \mid t_{D_i} \geq t_j, b_i, Z_{itj})$ at discrete time point $t_j \in \mathcal{C}$ to depend on the hypothetical complete measurement series $Y_i(\mathcal{C})$ through the unobserved covariate b_i . Specifically, we assume the model,

$$\log\{h_{itj}/(1 - h_{itj})\} = \varepsilon(Z_{it})\alpha + \delta\phi(b_i),$$

where $\varepsilon(Z_{it})$ is an appropriate function of fixed covariates Z_{it} with associated coefficient α . The parameter δ , assumed positive, is the slope associated to the transformed random effects $\phi(b_i)$. The function $\phi(\cdot)$ is a nondecreasing function such that $0 < \phi(b_i) < 1$, for all b_i , and attains the boundaries at the limits, that is $\lim_{b_i \rightarrow -\infty} \phi(b_i) = 0$ and $\lim_{b_i \rightarrow \infty} \phi(b_i) = 1$. Note that the function $\phi(\cdot)$ can suitably be modified when b_i has a finite support. We can reparameterise the function $\phi(\cdot)$ by writing $\phi(b_i) = (1 + e^{-\eta(b_i)})^{-1}$ for all b_i on the real line. The function $\phi(\cdot)$ imposes the following restrictions on $\eta(\cdot)$, $\lim_{b_i \rightarrow -\infty} \eta(b_i) = -\infty$ and $\lim_{b_i \rightarrow \infty} \eta(b_i) = 1$. Any such function $\phi(\cdot)$ that meets the conditions above defines a function $\eta(\cdot)$ and vice versa.

The idea of using a shared random effects model for modeling longitudinal outcome subject to dropouts not at random is not new (see for example, Wu and Carroll, 1988; Albert and Follmann, 2000 and Ten Have et al., 2002). This class of models is very attractive in the context of unbalanced longitudinal studies. Our extension of the shared random effects technique is the introduction of the function $\phi(\cdot)$ and restriction of the parameter δ to be positive. The function $\phi(\cdot)$ is assumed to be bounded, monotone, differentiable and invertible. These constraints endow the parameter δ with a meaningful interpretation as an

odds ratio for treatment assignment inequality (Rosenbaum, 2002), where the unobserved variable is constrained to the interval $[0, 1]$. Our constraint satisfies this condition, with the boundaries 0 and 1 attained at the limits. Here, the odds ratio of drop-out hazard between two subjects i and i' with same fixed covariate vectors, $Z_{it} = Z_{i't}$ is bounded as

$$\exp(-\delta) \leq \frac{h_{it}(1 - h_{i't})}{h_{i't}(1 - h_{it})} \leq \exp(\delta), \text{ with } \delta > 0.$$

In other words, for $\delta > 0$, $\exp(\delta)$ is a measure of the degree of departure from a study with a random dropout mechanism.

For the random effects model above, joint identifiability of the data and missingness models has not been established. A natural choice of sensitivity parameter is δ , which measures the extent of non-randomness of the dropout process. It is clear that when $\delta = 0$, the two processes can be separately identified and the dropouts and completers have the same distributional properties. Treating the fixed, possibly non-zero, value of δ as the true value, one may form a working likelihood under the random intercept model, assuming conditional independence of the outcomes and the missingness given random effects. The inference procedure we discuss below is valid assuming that $\psi_*(\delta)$ is identifiable conditionally on δ , regardless of the joint identifiability of the outcome model and the missingness model. Interestingly, even for the case of fixed δ , formal conditions for identifying $\psi_*(\delta)$ have yet to be developed.

Under mild conditions, for large δ the profile log-likelihood diverges and the maximum likelihood solution does not exist (proof shown in Supplementary Materials). That is,

$$\sum_{i=1}^{n_d} \log l_d^i(\psi | \delta) + \sum_{i=n_d+1}^n \log l_o^i(\psi | \delta) \rightarrow -\infty, \text{ as } \delta \rightarrow \infty,$$

where $n_d = \sum_i I(D_i \leq T)$ is the number of subjects dropping out prior to end of study, after a rearrangement of the data set. This supports the above constraint that δ is bounded. Hence, we assume without any loss of generality that δ lies in a compact set, that is $\delta \leq \Delta$ with $0 < \Delta < \infty$. In practice, the upper bound should be chosen so that the algorithm converges for all $\delta \leq \Delta$.

3.2 Analysis of psychiatric data

A good example of a longitudinal study with potentially non-random dropouts is the Fluvoxamine (a serotonin reuptake inhibitor) clinical trial. This is a multi-center non-comparative study, designed to reflect clinical practice closely with out-patients diagnosed with depression, obsessive-compulsive disorder or panic disorder. Accumulated experience in controlled trials has shown that Fluvoxamine is as effective as conventional tricyclic antidepressant drugs, and more effective than placebo in the treatment of depression (for a review, see Burton, 1991). However, many patients suffering from depression have concomitant morbidity associated with this condition, potentially leading to dropout. It was then decided to set up a post-marketing pharmaco-vigilance trial to study more accurately the profile of Fluvoxamine in ambulatory clinical psychiatric practice. A total of 315 patients with a diagnosis of either depression or obsessive-compulsive or panic disorder were enrolled in the study. All subjects were treated with Fluvoxamine in doses ranging from 100 to 300 mg/day and underwent clinical evaluations at baseline, 2, 4, 8 and 12 weeks. One primary endpoint comprised the side effects of the drug recorded on an ordinal

scale. Several patient's baseline characteristics such as sex, age, initial severity of the disease on a 1 to 7 scale, and duration of the mental illness were recorded. A full description of the study is given by Molenberghs and Lesaffre (1994), Lesaffre et al. (1996) and Kenward et al. (1994).

One key objective of the study was to assess the within-subject evolution of side effects over time, accounting for dropouts. For simplicity and ease of interpretation, our analysis of the within-subject evolution (captured by time effects) is based on a dichotomized version (presence/absence) of side effects, as opposed to ordinal outcomes, regardless of the baseline characteristics. A side effect occurs if new symptoms appear.

Figure 1 shows the empirical distribution of side effects over the four time points for the complete (full-sequence) cases and over the first three time points for dropout cases. Out of 315 patients, 224 patients had a full-sequence data resulting from the fact that 14 subjects were not observed after recruitment, 31, 26, and 18 patients dropped out, respectively, after the first, second and third visit and 2 patients had a non-monotone missing pattern. We ignore the 2 cases that had a non-monotone missing pattern as well as the 14 patients that dropped out before the first visit (at 2 months). Compared to the start of the study, a drastic reduction of side effects is depicted throughout the study for the completers compared to the dropout cases. Specifically, about 54% of patients who complete the study had some side effects at 2 months compared to 66% for subjects who did not complete the study. At 4 months, 44% of completers had some side effects compared to 62% for subjects who did not complete the study. And finally, at 8 months, 35% of completers had some side effects compared to 44% for subjects who did not complete the study. Clearly, the study non-completers are doing more poorly with respect to side effects. A naive analysis that ignores this selection process might lead us to conclude that the side effects drastically diminish over time in the studied population. A question emerges: How do side effects evolve over time within a patient when dropouts are accounted for?

As a preliminary analysis, we first fit a random subject effect model to the side effects response with only fixed linear time effects both to completers and non-completers. The measurement model is given by, $\text{pr}(Y_{it} = 1 | b_i, X_{it}) = \{1 + \exp(-\beta_0 - \beta_1 X_{it} - b_i)\}^{-1}$, where $\{Y_{it} = 1\}$ indicates the presence of some side effects and X_{it} represents the value of the time covariate for subject i at time point t . The results of the time slope estimates (standard errors) are $-.75 (.13)$ for the complete cases and $-.23 (.28)$ for the incomplete cases. The results of this analysis are consistent with those of the bar charts shown in Figure 1. In particular, the time effect on side effects appears larger (in magnitude) for the analysis based on the completers. Hence, subjects who dropped out prematurely from the study are likely to suffer more side-effects, although the estimate fails to reach statistical significance.

Following the sensitivity testing strategy described in section 2.3, we first fit the ignorable model to assess the effect of time on the side effects outcome. The result of $\hat{\beta}_1(0)$, the time slope estimate (standard error) is $-.76 (.12)$ using all available cases. The ignorable model clearly suggests a significant linear time effect (p-value $< .0001$).

We next attempted to simultaneously estimate δ and ψ via maximum likelihood. Unfortunately, the computations were unstable. Multiple starting values were tried. In some cases, the algorithm diverged, while in cases where it did converge, multiple local maxima were obtained. This suggests that the model is at best weakly identified on the psychiatric data.

In order to conduct inferences which do not require joint identifiability of the data and missingness models, we evaluate the hypothesis $\inf_{\delta} |\beta_{*,1}(\delta)| > 0$ using the infimum test under the shared random effects with a dropout model given by; $\log\{h_{it}/(1 - h_{it})\} = \alpha_0 +$

$\alpha_1 X_{it} + \delta \phi(b_i)$. Our sensitivity analysis was restricted to the range $[0, 80]$ as values of δ above 80 yields unstable estimates. We perform the infimum test based on 1000 bootstrap samples, using the statistic $T = \inf_{\delta} \{ \hat{\beta}_1(\delta)' \hat{\Sigma}(\delta)^{-1} \hat{\beta}_1(\delta) \}$. The observed value of the infimum test statistic is $T^{obs} = 40.90$, which is highly significant at 1% level (p-value $< .0001$) (see the histogram of bootstrap test samples in Figure 2). This test suggests that $\inf_{\delta} |\beta_{*,1}(\delta)|$ is greater than zero and it can then be concluded that the true time effect measured by $|\beta_{*,1}(\delta_0)|$ is greater than zero.

One might criticize the choice of the upper bound $\Delta = 80$ as being scientifically unreasonable. We believe that this choice reflects an extreme scenario, which may be of scientific interest, albeit one which is a priori unlikely. The main concern is that contemplating such scenarios leads to a null hypothesis which reflects an overly conservative worst case scenario. Interestingly, in this example, the evidence for the covariate effect is rather strong, so that the infimum test rejects over this wide interval. Moreover, in general, considering such a wide range for δ enables one to construct confidence regions for δ for which $|\beta_{*,1}| > 0$, which is practically useful. That is, one can determine all values of δ for which a non-zero covariate effect exists, accounting for simultaneous inference across many δ .

To illustrate how the magnitude of the covariate effect changes as the sensitivity parameter is varied, we also computed simultaneous 95% confidence bands for $\beta_{*,1}(\delta)$ for $\delta \leq 80$ (see Figure 3). All simultaneous confidence intervals exclude 0, which is in agreement with the infimum test above. The largest time effect is achieved when $\delta \approx 15$.

This example illustrates that it may be useful to conceptualize a sensitivity analysis in terms of testing the smallest covariate effect over the range of the sensitivity parameter when the true effect is not identifiable from the observed data. The approach is especially useful in high-risk situations where a worst case analysis may be helpful. For the Fluvoxamine data set, the linear time effect exits regardless of the sensitivity parameter. In addition, we found that the estimate of the time effects ranges between -1.15 and $-.74$ given values of the sensitivity parameter. Hence, the estimated odds for showing some side effects can be multiplied by a factor as small as $.32$ and as much as $.48$ for each unit increase in time. The 95% confidence bounds run from roughly -1.50 (smallest value on lower bound across all δ) to $-.50$ (largest value on upper bound across all δ). This could be viewed as a conservative 95% confidence interval for the unknown $\beta_1 = \beta_{*,1}(\delta_0)$.

In this example, the global analysis is essential since the estimate of the parameter being tested does not increase or decrease monotonically as the sensitivity parameter is increased or decreased. An explanation is that the joint measurement-missingness model is highly nonlinear, so that there may be a complex relationship between $\beta_*(\delta)$ and δ , which may not be monotone in δ . Under monotonicity, it is only necessary to evaluate $\beta_*(\delta)$ at the lower and upper bounds of the sensitivity parameter space in order to test the null hypothesis. With non-monotone behavior, such tests may yield misleading results.

3.3 Simulations

We report results of a small simulation study evaluating the performance of the infimum test under the null when the dropout process is related to unobserved responses. The simulations were conducted so as to roughly approximate data from the Fluvoxamine study. In each Monte Carlo iteration, we simulated a sample of n subjects with four potential measurement time points ($T = 4$), using the outcome model $\text{pr}(Y_{it} = 1 | b_i, X_{it}) = \{1 + \exp(-\beta_0 - \beta_1 X_{it} - b_i)\}^{-1}$, where $b_i \sim N(0, 2)$ and $X_{it} = t$ takes values in $\{1, 2, 3, 4\}$. The dropout observations were generated using the dropout hazard model $\log\{h_{it}/(1 - h_{it})\} = \alpha_0 + \alpha_1 X_{it} + \delta_0 \phi(b_i)$. To keep the simulation simple, we set $\alpha_1 = 0$ to constrain the instantaneous dropout probability to be

time independent and $\beta_1 = 0$, so that the null is satisfied. We produced 20 – 60% dropout using both an ignorable ($\beta_0 = 0$) and a non-ignorable ($\beta_0 = 3$) missing data models. This process was repeated for 1000 Monte Carlo replications. The infimum tests were computed using 1000 bootstrap samples for $\delta \leq 10$.

Table 1 shows the results of this simulation study for sample sizes equal 25, 50, 100, and 250 and nominal test levels equal to .01, .05, and .1. Asymptotic standard errors (as the number of Monte Carlo iterations tends to infinity) are reported in the last row of the table. Overall, the tests perform well, with the bootstrap distribution of the test providing a reasonable approximation to the nominal level even in fairly small samples ($n = 25$). As expected, for increasing sample sizes, the empirical rejection rates agree more closely with the true significance levels under the null.

4. Discussion

In this paper, we have shown how a statistical hypothesis can be evaluated using a joint parametric model for the measurement and missing data processes, when some model characteristics are non-identifiable. We proposed an approach in which hypothesized parameters are viewed as a functional of a non-identifiable sensitivity parameter. A conservative testing procedure was employed based on an infimum statistic over the range of the sensitivity parameter. Similar approaches based on confidence bounds have been studied for global sensitivity analyses under qualitative restrictions. Our approach differs in that we work with parametric models and rigorously delineate the full spectrum of inferences consistent with the model assumptions. Previous work with parametric models has not addressed key issues inherent in such formulation, including multiple testing and model misspecification. Our methodology enables the construction of simultaneous confidence bands which pinpoint those values of the sensitivity parameter at which significance is achieved, as well as a comparison of the magnitude of the effects across the sensitivity parameter.

Obtaining the sensitivity parameter region over which a covariate effect is significant may be particularly important in scenarios where the effect does not increase or decrease monotonically as the sensitivity parameters are increased or decreased. In practice, missing data models are often quite complicated and it may not be clear whether monotonicity holds. In these set-ups, a global sensitivity analysis is the most conservative data analytic strategy. For the Fluvoxamine example, the infimum test shows that the linear time effect exists, regardless of the sensitivity parameter, while the confidence bands demonstrate that the magnitude of the effect is non-monotone. These findings are extremely important for the drug company which may be interested in knowing how the side effects evolve within a subject as a function of the drop-out mechanism.

Although the method is presented in the context of likelihood based analyses, with a focus on generalized linear mixed effects models for binary outcomes in time with non-ignorable dropouts, the method can easily be applied to other situations where any component of the model is non-identifiable from observed data and non-likelihood based methods of analysis are available, for example, generalized estimating equations. Further research is needed to show how to choose the sensitivity parameter in other model settings. The choice of the sensitivity parameter from a statistical perspective is non-unique, so long as identifiability is induced in the appropriate sense. However, a careful choice may lead to simplifications in the computations (Kenward et al., 2001). The subject matter is also instrumental in guiding the investigator in the choice of the sensitivity parameter.

Additional work is needed to show how to apply the global test in situations where more than one dimension of the model may be non-identifiable. For example, for our working random effects model, multiple sensitivity parameters could be considered, if it was deemed necessary for identification purposes. The parameters in the shared random effects distribution might be treated as sensitivity parameters along with the parameter δ , yielding a multidimensional problem. The infimum test can then be constructed over a class of distributions for the random effects along with values of δ . The computational burden of such tests will increase with the dimension of the sensitivity parameter and be burdensome in some setups. The computational issues require careful study. Another issue is that we restrict δ to be bounded. It would be useful to permit the sensitivity parameters to be unbounded. While this is conceptually straightforward, it poses certain theoretical challenges which are not easily addressed. This is a topic for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank Solvay Pharma B.V. for permission to use the data from the Fluvoxamine study. They also wish to thank Geert Molenberghs and the anonymous reviewer for their helpful comments. This work was supported by the NCI/NIH K-award, 1K01 CA131259 (David Todem).

REFERENCES

- Albert PS, Follmann DA. Modeling repeated count data subject to informative dropout. *Biometrics*. 2000; 56:667–677. [PubMed: 10985201]
- Burton SW. A review of the fluvoxamine and its uses in depression. *International Clinical Psychopharmacology*. 1991; 6 Supplement 3:1–17. [PubMed: 1806631]
- Copas J, Eguchi S. Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B: Methodological*. 2001; 63:871–895.
- Efron, B.; Tibshirani, R. *An introduction to the bootstrap*. London: Chapman and Hall; 1993.
- Horowitz JL, Manski CF. Identification and estimation of statistical functionals from incomplete data. *Journal of Econometrics*. 2006; 132:445–469.
- Kenward MG, Lesaffre E, Molenberghs G. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*. 1994; 50:945–953. [PubMed: 7787007]
- Lesaffre E, Molenberghs G, Dewulf L. Effects of dropouts in a longitudinal study: An application of a repeated ordinal model. *Statistics in Medicine*. 1996; 15:1123–1141. [PubMed: 8804143]
- Manski, C. *Partial identification of probability distributions*. New York: Springer; 2003.
- McCullagh P, Tibshirani R. A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society B*. 1990; 52:325–344.
- Molenberghs G, Lesaffre E. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*. 1994; 89:633–644.
- Rosenbaum, PR. *Observational Studies*. New York: Springer; 2002.
- Scharfstein DO, Manski CF, Anthony JC. On the construction of bounds in prospective studies with missing ordinal outcomes: Application to the good behavior game trial. *Biometrics*. 2004; 60:154–164. [PubMed: 15032785]
- Shepherd BE, Gilbert PB, Jemimai Y, Rotnitzky A. Sensitivity analysis comparing outcomes in a subset selected post randomization, conditional on covariates, with application to hiv trials. *Biometrics*. 2006; 62:332–342. [PubMed: 16918897]
- Ten Have TR, Reboussin BA, Miller ME, Kunselman A. Mixed effects logistic regression models for multiple longitudinal binary functional limitation responses with informative drop-out and confounding by baseline outcomes. *Biometrics*. 2002; 58:137–144. [PubMed: 11890309]

- Todem D, Kim K, Lesaffre E. A sensitivity approach to modeling longitudinal bivariate ordered data subject to informative dropouts. *Health services and outcomes research methodology*. 2006 *in press*.
- Troxel A, Ma G, Heitjan D. An index of local sensitivity to nonignorability. *Statistica Sinica*. 2002; 14:1221–1237.
- van der Vaart, AW.; Wellner, JA. *Weak convergence and empirical processes*. New York: Springer; 2000.
- Vansteelandt S, Goetghebeur E, Kenward MG, Molenberghs G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*. 2006; 16:953–979.
- Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward M. Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics*. 2001; 57:7–14. [PubMed: 11252620]
- White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50:141–161.
- Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988; 44:175–188.

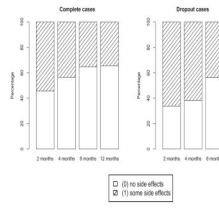


Figure 1.
Evolution of observed side effects for all complete cases over observed time points

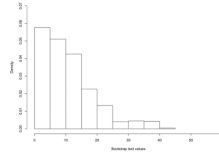


Figure 2. Histogram of the 1000 bootstrap sample test statistics generated under the null hypothesis

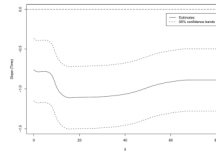


Figure 3. Estimate of time slope for fixed values of the sensitivity parameter δ and corresponding 95% simultaneous confidence bands using the side effects data from the Fluvoxamine study

Table 1

Monte carlo empirical rejection rates in simulations study

True dropout model		Sample size				
Mechanism	% dropout	(α_0, δ_0)	n=25	n=50	n=100	n=250
			Nominal test level	Nominal test level	Nominal test level	Nominal test level
Random	20	(2.55, 0)	.01 .05 .10	.01 .05 .10	.01 .05 .10	.01 .05 .10
			.080 .155 .006	.081 .158 .009	.072 .129 .007	.064 .124
dropouts	40	(1.67, 0)	.013 .065 .143	.014 .073 .148	.017 .060 .127	.013 .055 .119
			.004 .067 .150	.001 .043 .112	.004 .050 .118	.012 .057 .114
Non-random	20	(1.34, 3)	.014 .071 .152	.014 .073 .150	.011 .063 .138	.014 .063 .118
dropouts	40	(.34, 3)	.012 .098 .178	.006 .054 .133	.006 .067 .131	.010 .074 .122
	60	(-.52, 3)	.007 .059 .130	.008 .037 .099	.012 .064 .127	.015 .058 .125
Monte carlo s.e.			.003 .007 .009	.003 .007 .009	.003 .007 .009	.003 .007 .009