ADDITIONAL REFERENCES

Belin, T. R., and Rolph, J. E. (1994), "Can We Reach Consensus on Census Adjustment?" with discussion, *Statistical Science*, 9, 486–508.

Copas, J. B., and Li, H. G. (1997), "Inference for Non-Random Samples," *Journal of the Royal Statistical Society*, Ser. B, 59, 55–77.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, (2nd. ed.), New York: Wiley.

Freedman, D., and Wachter, K. (1994), "Heterogeneity and Census Adjustment for the Post-Censal Base" with discussion, *Statistical Science*, 9, 476–485, 527–537.

Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.

Shaffer, J. P. (1992), *The Role of Models in Nonexperimental Social Science: Two Debates*, Washington, DC: American Educational Research Association and American Statistical Association.

Comment

Jianqing FAN and Chunming ZHANG

Adjusting for nonignorable drop-out is an interesting and challenging topic in statistics that has significant impact on statistical decision and policy making. Because results can be altered significantly by different assumptions on drop-out time, any conclusions based on adjustments should be drawn with care. Assumptions should be checked rigorously. Adjusting for nonignorable responses is arguably one of the most debatable subjects in statistics. We welcome the opportunity to make a few comments.

Scharfstein, Rotnitzky, and Robins are to be congratulated for this excellent article on adjusting for nonignorable drop-out using semiparametric nonresponse models. It elegantly lays out a semiparametric framework and offers useful tools for sensitivity analyses. It provides insightful and convenient models for assessing nonresponse biases. In contrast with the missing-at-random assumption, this article allows one to explore how conclusions can be affected if data are not missing at random.

1. CAN DROP-OUT BIASES BE ADJUSTED FAIRLY?

This is a fair question one naturally asks. The answer depends on how well one can relate a response variable with observable variables. To account for possible nonresponse biases, an assumption on how subjects dropped out during a study must be made. Without such a critical assumption, it is not possible to assess the drop-out biases with reasonably good accuracy. Indeed, to some extent, adjustments depend purely on the assumptions on the drop-out process. Different assumptions can result in completely different conclusions. This leads to final conclusions that are inevitably subjective and disputable. The authors hinted repeatedly that subject matter experts can be consulted on the choice of models for the drop-out process. This is indeed very helpful for ruling out some unrealistic models. However, chances are that subject matter experts themselves cannot be certain why drop-out occurs during a study and hence might not feel comfortable with adjustments if such adjustments cannot be validated scientifically. Things can be worse when political or emotional factors get involved in the selection of adjustment methods. These kinds of "adjustment biases" can be very severe in extreme cases. Thus model validation is very important for an adjustment method. A question then arises whether one should adjust at all for drop-out biases if no reliable method is available for modeling the drop-out time.

There are infinitely many possibilities for modeling drop-out risk. To account for the nonignorable biases, it is assumed that the drop-out risk follows the Cox proportional hazards model. This assumption is the most important and most arguable one in the article. The adjustments are basically reflections of this critical model assumption. One concern is that the model is not driven by any physical law, and is not derived from conceivable intuitions. There is always the risk that the model is misspecified. This possibility cannot be rescued simply by considering a few other classes of models, such as additive hazards models or other parametric forms $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$. The sensitivity to model specifications make adjustment procedures debatable.

The article offers few clues as to why the Cox proportional hazards model is chosen for modeling the risk of drop-out. It would be very helpful if the authors elucidated it. One speculation is that it is a convenient model for handling censored data. But there are also some costs entailed in using this model. One needs to evaluate the baseline hazard function at study termination time T. The value T can be in the tail region of the distribution of drop-out time Q. This tail probability usually cannot be estimated reliably. Further, the authors assume that the drop-out time Q is continuously observable. If Q is observed only at a few prescheduled time periods of clinical evaluation, then the variable Q is censored. The entire statistical analysis can be far more complicated than the current setting.

That the sensitivity of adjustments depends critically on model assumption is convincingly demonstrated by the authors in their Figure 1. With a small change of assumption on drop-out time done by varying a sensitivity parameter,

Jianqing Fan is Professor, Department of Statistics, University of California, Los Angeles, CA 90095. Chunming Zhang is a graduate student, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. This work was partially supported by National Science Foundation grant DMS-9803200.

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

mean CD4 counts can range from very small to very large. This makes comparisons of several treatments very difficult. The authors compared two treatment arms for different combinations of sensitivity parameters, based on the same family of models on the drop-out time (see the authors' Fig. 2). Although this method is very useful, it is conceivably possible that the drop-out risk for one treatment group follows a family of stochastic models but follows a completely different one for the other treatment group. Thus as long as there are no data available for validation of model assumptions, there are always uncertainties on an adjustment method.

2. VALIDATION OF ASSUMPTIONS ON DROP-OUT TIME

Validation of model assumptions on drop-out time is particularly important for adjusting for nonignorable drop-out. It reduces the chance of making erroneous adjustments. It is, however, hard to check all aspects of model assumptions. To attenuate the difficulty, the authors correctly pointed out that one should consult with subjectmatter experts. One can also use other information collected during the course of a study to validate certain aspects of model assumptions.

There are several possible methods to accommodate side information. Take clinical trial ACTG 175 AIDS, for instance. One can use available CD4 measurements at earlier weeks to predict missing measurements at week 56. The results can then be used to calibrate the sensitivity analysis for the data at week 56. This will validate the extent to which the two adjustment methods are consistent.

A second possibility is to compute implied (model-based) missing probability for different choices of the sensitivity parameter α_0. Despite α_0 unidentifiable, this checking verifies how well unknown parameters and functions were estimated. Under model (1), the missing probability is given by

$$P(\Delta = 0) = E\{1 - S(T|\bar{\mathbf{V}}(T), Y)\},$$

where $S(t|\bar{\mathbf{V}}(T), Y) = \exp(-\Lambda_0(t|\bar{\mathbf{V}}(t)) \exp(\alpha_0 Y))$. As in Section 3 of the article, if $\bar{\mathbf{V}}(t)$ is a time-homogeneous discrete covariate, then the distribution of Y for each level of $\mathbf{V}$ can be estimated. Hence an estimate of the implied missing probability can be obtained. If the situation is more complicated than the foregoing simple setting, then one can use the crude estimate

$$P(\Delta = 0) \approx n^{-1} \sum_{i=1}^{n} \{1 - \hat{S}(T_i|\bar{\mathbf{V}}_i(T), \hat{Y}_i)\},$$

where for missing cases, $\bar{\mathbf{V}}_i(T)$ is $\bar{\mathbf{V}}_i(Q_i)$ and $\hat{Y}_i$ is the imputed response such as estimated population mean. The implied missing probability can be estimated more carefully than what is outlined. An estimated missing probability under model (1) that is excessively larger or smaller than the sample proportion of missing data provides evidence that the drop-out model is inadequate.

A third possibility of model validation is to use the same drop-out model to analyze the mean CD4 counts at an earlier time, such as week 20, 32, or 44. For the ACTG 175 data, there are on average about 15.6%, 21.0%, and 25.7% of drop-out. A low percentage of drop-out means that the mean CD4 counts can be estimated more reliably even without adjustments (missing at random). Applying the same sensitivity analysis techniques to the data collected at these earlier weeks, if a drop-out risk model is right and the range of sensitivity parameter is reasonable, then the estimated population mean should be less sensitive than at week 56. Should the results contradict with this intuition, it would be odd to accept the assumption that the drop-out risk is modeled correctly. This gives us an idea to verify whether the model is reasonable and the sensitivity parameter is in a good range.

3. CHOICE OF SENSITIVITY PARAMETERS

We wholeheartedly endorse the notion that the sensitivity parameter α_0 should be varied to see how sensitive conclusions depend on this parameter. It appears unclear, however, what the scale of α_0 is and what range of α_0 should be used. An extreme value of α_0 can not only make resulting estimates meaningless, but also cause some technical problems and numerical instability.

Different values of α_0 entail different implied missing probabilities under a drop-out model. A possible choice of the range of α_0 is to make the model-based missing probabilities consistent with empirical (observed) missing probability. An implied missing probability that is much too large or too small provides evidence that the choice of α_0 does not fit available data. See Section 2 of this discussion for a simple method of calculating the model-based missing probability.

Another possibility for choosing a range of a sensitivity parameter is to apply sensitivity analysis to the data collected at an earlier stage of a longitudinal study where the percentage of drop-out is smaller. The sensitivity parameter can be chosen so that the resulting estimate is reasonably close to the estimate deriving directly from the data collected at that time point.

4. OTHER APPROACHES TO COMPARING TWO TREATMENTS

Population mean is known to be not robust. It is not surprising that estimates of a population mean depend sensitively on the assumptions of how data were missing. Given the risks of model misspecification and the difficulty of model validation, it seems preferable to using a more robust functional of an underlying distribution, such as median, trimmed mean, or the distribution function itself. As reported by the authors in Section 7.1, such robust functionals are less sensitive to underlying models on drop-out time. Hence they are more objective and less disputable for assessing the efficacy of treatments.

When two treatments are compared, we are interested not only in whether two estimates of population means are significantly different, but also in how different the treatment effects are. An illuminating method is to compare the two distributions of the response variables. For estimation

of a cumulative distribution $F(y)$ under the authors' framework, one can solve an equation similar to (3) for each given y. Namely, one could estimate $F(y)$ by solving the equation

$$\sum_{i=1}^{n} h(\mathbf{O}_i; F, \Lambda_0; b) = 0$$

with

$$h(\mathbf{O}; F, \Lambda_0; b) = \frac{\Delta}{\pi(\bar{\mathbf{V}}(T), Y)} \, [I(Y \le y) - F - E\{(1-\Delta) \times b(\bar{\mathbf{V}}(Q), Q; F) | \bar{\mathbf{V}}(T), Y)\}] + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q; F).$$

The distributions can be informatively and visually compared by computing their corresponding kernel density estimates (see, e.g., Fan and Gijbels 1996; Silverman 1986). For a given bandwidth h and a kernel function $K(\cdot)$ (e.g., a symmetric probability density function), the estimates can be obtained via smoothing the estimated distribution functions as follows:

$$\hat{f}(y) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{y-t}{h}\right) d\hat{F}(t).$$

Using the estimated densities, one can compare how the two treatments differ from each other, not only in population mean, but also in dispersion and other important functionals (see Fan and Gijbels 1996, p. 49).

As an illustration, we plot the estimated densities for the four treatment arms, using a boundary corrected kernel density estimator. Subjects are assumed to be missing at random for simplicity of computation. Figure 1 indicates that all four treatment arms have similar-shaped densities of CD4 counts. But the treatment using AZT has lower CD4 counts, and the other three treatment arms perform very similarly.

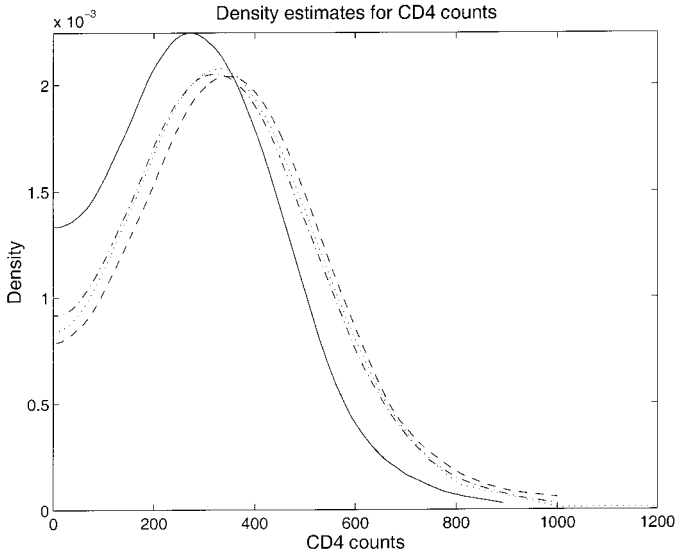

Figure 1. Kernel Density Estimates for CD4 Counts for Four Separate Treatment Arms, Assuming that Subjects Were Missing at Random. ——, AZT; – – –, AZT+ddI; · · ·, AZT+ddC; · - · - · -, ddI.

5. UNDERSTANDING ESTIMATING EQUATIONS

We now give a simple derivation for estimators given in the article and relate them with the existing techniques in the censored regression literature. Assume that $b(\bar{\mathbf{V}}(T), t, \mu)$ is independent of μ. Then (3) gives the solution

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i^*,$$

where

$$Y^* = \pi(\bar{\mathbf{V}}(T), Y)^{-1}[Y - E\{(1-\Delta) b(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y)\}] \quad \text{if not missing}$$

$$= b(\bar{\mathbf{V}}(Q), Q) \quad \text{if missing.}$$

This estimator looks complicated but can be simply derived as follows. One naturally uses the available information $(\bar{\mathbf{V}}(Q), Q)$ to impute missing data, resulting in adjustment $b_0(\bar{\mathbf{V}}(Q), Q)$, for a given function b_0. Correspondingly, one adjusts a nonmissing case by $b_1(Y, \bar{\mathbf{V}}(T))$, using all collected information. Then the data after adjustments become

$$Y^* = \Delta b_1(Y, \bar{\mathbf{V}}(T)) + (1-\Delta) b_0(\bar{\mathbf{V}}(Q), Q).$$

For the sample average of the adjusted data to be unbiased, it is required that

$$\begin{aligned} EY &= EY^* \\ &= E[\pi(\bar{\mathbf{V}}(T), Y) b_1(Y, \bar{\mathbf{V}}(T)) \\ &\quad + E\{(1-\Delta) b_0(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y\}]. \end{aligned}$$

This equation is obviously satisfied if

$$Y = \pi(\bar{\mathbf{V}}(T), Y) b_1(Y, \bar{\mathbf{V}}(T)) + E\{(1-\Delta) b_0(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y\},$$

or, equivalently,

$$b_1(Y, \bar{\mathbf{V}}(T)) = \pi(\bar{\mathbf{V}}(T), Y)^{-1} \times [Y - E\{(1-\Delta) b_0(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y\}].$$

This yields exactly the same procedure as that of the authors.

The idea of the foregoing derivation appeared already in the censored regression literature (see, e.g., Fan and Gijbels 1994; Zheng 1987). To get the best estimator after adjustments, Fan and Gijbels (1994) argued that the function b_0 should be chosen such that $\text{var}(Y^*)$ is minimized, among a class of possible adjustment functions under consideration. An intuitive and appealing transformation function is to take

$$b_1(Y, \bar{\mathbf{V}}(T)) = Y, \quad b_0(\bar{\mathbf{V}}(Q), Q) = E(Y | Q, \bar{\mathbf{V}}(Q), \Delta = 0),$$

or, equivalently,

$$Y^* = E(Y | Q, \Delta, \Delta Y, \bar{\mathbf{V}}(Q)).$$

This pair of adjustments of course satisfy the requirement $EY = EY^*$ and are the best restoration in the sense that it is closest to the original data Y. This transformation is related to the Buckley–James (Buckley and James 1979) method in the censored regression setting.

The foregoing adjustment functions depend on unknown parameters. To implement the idea in practice, one needs to estimate these parameters using observable data. A parametric model and a semiparametric model are proposed by Scharfstein, Rotnitzky, and Robins for the purpose of estimating these unknown parameters. The resulting estimator is the sample mean of the adjusted data. The idea is also applicable to other linear functionals, such as the cumulative distribution function at a point. With the estimated cumulative distribution function, all other functionals can easily be estimated via a plug-in method.

ADDITIONAL REFERENCES

Buckley, J., and James, I. R. (1979), "Linear Regression With Censored Data," *Biometrika*, 66, 429–436.

Fan, J., and Gijbels, I. (1994), "Censored Regression: Nonparametric Techniques and Their Applications," *Journal of American Statistical Association*, 89, 560–570.

——— (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Zheng, Z. (1988), "Strong Consistency of Nonparametric Regression Estimates With Censored Data," *Journal of Mathematical Research and Exposition*, 3, 231–241.

Comment

Mark VAN DER LAAN

Let me first compliment the authors on their two very nice articles on sensitivity analysis. The article under discussion applies a nonparametric sensitivity analysis toward a particular type of censored-data model of practical interest. The methodology is completely general and thus can be applied to any censored-data structure of a full-data random variable X; that is, when the observed data O can be defined as $O = \Phi(X, C)$ for some given mapping Φ and censoring variable C. A model of such a censored-data structure is typically built up by a model for the full-data random variable X and a model for the conditional distribution C, given X, referred to as the censoring mechanism. To make estimation of the full-data distribution or parameters thereof tractable, one often assumes that the censoring mechanism satisfies coarsening at random (CAR) as originally formulated by Heitjan and Rubin (1991) and generalized by Jacobsen and Keiding (1995) and then by Gill, van der Laan, and Robins (1997). Under CAR, the likelihood factorizes into a likelihood for the full-data distribution and a likelihood for the censoring mechanism, so that maximum likelihood estimation of the full-data distribution can ignore the censoring likelihood. Intuitive understanding of the CAR assumption is in general very hard, but in monotone-censored data structures (i.e., one observes an increasing sigma field $\mathcal{F}_t$ over time t up till the minimum of C and the point at which the full data are completely observed) it has a very appealing and easily understandable interpretation. In these monotone-censored data problems, the censoring mechanism satisfies CAR if and only if the hazard of C at t, given the full data X, is only a function of the data $\mathcal{F}_t$ that one has available at time t. The particular censored-data structure handled by the authors is monotone. Thus to have CAR, one wants to collect over time any variables that might be used in a "decision" to censor a subject. An estimator that is consistent under CAR will extrapolate a subject censored at time t by using uncensored subjects who have the same observed history up until time t. If one feels that in a particular application uncensored and censored subjects at time t with the same observed history might not be exchangeable, then the CAR assumption will need to be tested and/or a sensitivity analysis will be appropriate.

In nonparametric CAR censored-data models in which the full-data distribution is unspecified and the censoring mechanism is only known to satisfy CAR, one cannot test whether the censoring mechanism satisfies coarsening at random, because the model is already nonparametric. However, to get a good sense about deviations from CAR, one could assume a semiparametric CAR model for the censoring mechanism and extend this with a one-dimensional parameter α measuring a deviation from CAR. Now one could try to estimate all unknown parameters including α, which will then provide a test of CAR given the assumed model. Under the latter extension of the CAR model, the likelihood will not factorize anymore. As a consequence, maximum likelihood estimation is typically less attractive and also an estimating equation approach is harder (see Rotnitzky, Robins, and Scharfstein 1998).

In a nonparametric CAR model, one can still extend the CAR model for the censoring mechanism with a one-dimensional parameter α measuring the deviation from

Mark van der Laan is Associate Professor, Division of Biostatistics, School of Public Health, University of California, Berkeley, CA 94720.

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

CAR, but now α will no longer be identifiable. The authors' proposed methodology is developed for such nonparametric CAR models. They develop estimators and confidence intervals for a range of given α's, provide the implied sensitivity plots, and mention that "when possible, one should try to leave the laws of X and Λ_0 completely unspecified and perform a sensitivity analysis." Thus the authors prefer such a nonparametric sensitivity analysis assuming a more parametric censoring or full-data model and estimating α accordingly, because it is their belief that a secure scientific basis for model assumptions is rarely available. However, they do want to rely on scientific experts to obtain a plausible range of α values as needed in the nonparametric sensitivity analysis.

I find the nonparametric sensitivity analysis attractive because it determines the sensitivity toward deviations from CAR under minimal assumptions. But relying on the needed scientific knowledge to obtain a plausible range of α values might be more problematic than making certain model assumptions on either the full data or the censoring law. The other extreme is to assume a parametric model for the censoring mechanism so that α is identifiable from the data, as done by Rotnitzky et al. (1998). There is no win-win situation, because the latter method might yield biased inference due to misspecification of the censoring mechanism. I propose using a lower-dimensional model for the censoring mechanism to data adaptively obtain a plausible range of α values for the nonparametric sensitivity analysis. This plausible range will be correct if the lower-dimensional model is correct and typically will be overly optimistic when the lower-dimensional model is wrong, but still provides something to work with. To be concrete, I consider the ACTG 175 trial as analyzed in the article. I argue that to make this nonparametric sensitivity analysis approach practical, one will need to provide data-adaptive ways to provide such a plausible range of α values.

In each of the four treatment arms of the ACTG 175 trial, each subject's CD4 count process is observed up till the minimum of 56 weeks and the subject's drop out time Q. In addition, one observes baseline covariates such as the IV drug user status of the subject. The observed data structure is a monotone censored-data structure in the sense that the amount of information one observes on a subject increases over time; formally, the sigma fields $\mathcal{F}_Q$ generated by $(Q \wedge 56, \bar{\mathbf{V}}(Q \wedge 56))$ are increasing in Q. If one is willing to assume coarsening at random—that is, that $\Lambda_Q(dt|\bar{\mathbf{V}}(56)) = \Lambda_Q(dt|\bar{\mathbf{V}}(t))$ for $t < Q$—then the results in the appendixes of Robins (1993) and Robins and Rotnitzky (1992) provide closed-form locally efficient estimators of, for example, the distribution function of $Y = \mathrm{CD4}(56)$. These estimators rely on an estimator of the drop-out mechanism Λ_Q.

Would this analysis be appropriate? In other words, is it reasonable to assume that the decision of a subject to drop out of the study at time t only depends on the subject's past CD4 count history and possibly other measured variables? The latter assumption corresponds in the authors' notation with model $A(0)$. Because this assumption does not strike me as unreasonable, I would have been satisfied with a locally efficient data analysis assuming a Cox proportional hazards model for the drop-out mechanism with time-dependent covariates including subjects' CD4 counts. The authors' nonparametric sensitivity analysis assumes for a fixed known α that

$$\Lambda_Q(t|\bar{\mathbf{V}}(\tau)) = \Lambda_0(t, \bar{\mathbf{V}}(t)) \exp(\alpha Y)$$

for some unspecified Λ_0, which corresponds with assuming model $A(\alpha)$, and they estimate the distribution of Y under this model $A(\alpha)$. This analysis is repeated for a plausible range of values α. Because it is already very hard to understand why CAR would be violated in this application, it will be much harder to determine a plausible range of α values. However, it might be easy to reason that α's smaller than a given ε cannot be excluded as a possibility, so that an extremely sensitive sensitivity plot would send a warning that I would have not known of without the sensitivity analysis.

Each of the models $A(\alpha)$ identifies the distribution of Y and is nonparametric. The advantage of nonparametrically identified models is that the conclusions of a sensitivity analysis are not affected by misspecification of the observed data model. On the other hand, because all models $A(\alpha)$ for various α are nonparametric, the data cannot distinguish between different $A(\alpha)$'s. Thus in each application one needs a certain type of expert who can provide a plausible range of α values. This requires experts who can tell to what degree a person with a given covariate history up to time t bases his or her decision to drop out at time t on his or her future CD4 count value. One might wonder if such experts exist.

As pointed out by the authors, one can choose different types of sensitivity models for the drop-out mechanism, and it makes sense to choose the one that is easiest to interpret with regard to α. However, whatever model one selects, the experts' minimal task will still be to determine with respect to some measure to what degree subjects' drop-out time behavior deviates from CAR. In addition, because the conclusions (the sensitivity plot) depend on the choice of sensitivity model, in principle even the choice of sensitivity model should be determined by a so-called expert. The true sensitivity parameter is not α, but in fact is the whole function $r(Y, \bar{\mathbf{V}}(t), \alpha)$, which makes the task of the expert even harder.

Finally, because the estimators of $\mu(\alpha)$ are IPCW estimators, even when one succeeds in determining a set of plausible α values, the estimators might break down for values of α within this range. In that case one would need to conclude for these plausible values of α that $\mu(\alpha)$ is not identifiable (for the given sample size).

Thus in many applications it is likely that the desired expert knowledge is not available. Then the data analyst who is concerned about the censoring mechanism not being ignorable will need other tools to get a plausible range of α values in another manner. Assuming, as the authors need to do to fight the curse of dimensionality, a Cox proportional

hazards model for Λ_0—that is,

$$\Lambda_0(dt, \bar{\mathbf{V}}(t)) = \Lambda_0(dt)\exp(\boldsymbol{\gamma} Z(t)), \tag{1}$$

with $Z(t)$ a function of $\bar{\mathbf{V}}(t)$)—the observed data model with α known is not nonparametric anymore, so that for, a sufficiently low-dimensional $Z(t), \alpha$ will be identifiable from the data. For example, in the ACTG 175 trial it would be of interest to estimate α when assuming model 1. Because α is only one-dimensional, the authors' argument that α might still be extremely hard to identify in the ACTG 175 trial needs to be proved and very well might not hold.

Consider this latter model with α being a parameter. The class of estimating equations for $(\mu, \boldsymbol{\gamma}, \alpha)$ can be derived as in the Appendix of the article by determining the orthogonal complement of the nuisance tangent space of $(\mu, \boldsymbol{\gamma}, \alpha)$. In fact, the authors already determined the orthogonal complement of the nuisance tangent space of $(\mu, \boldsymbol{\gamma})$ in the model with α known. Thus one simply subtracts from each element in this latter space the projection on the score of α to obtain the class of estimating equations for $(\mu, \boldsymbol{\gamma})$. However, this does not yield yet the estimating equation for α, but determining the orthogonal complement of the nuisance tangent space of α will not be a harder task than the work the authors have already carried out. This results in a set of estimating equations for $(\mu, \boldsymbol{\gamma}, \alpha)$ that still requires estimating of the nuisance parameter Λ_0. However, for a given $(\boldsymbol{\gamma}, \alpha)$, we can still estimate the baseline hazard Λ_0 as in the article. Thus this gives a complete set of estimating equations for all unknown parameters $(\mu, \boldsymbol{\gamma}, \alpha, \Lambda_0)$. In this manner one can obtain an confidence interval for the true α, and there would then be no need for expert knowledge.

This approach is against the authors' philosophy, because they really want to aim at a nonparametric model for Λ_0. The only reason that they select in their data analysis a Cox model for Λ_0 is that it is needed to make estimators available at all, but their goal is to choose the model as nonparametric as sample size allows, though this philosophy of selecting as nonparametric model as sample size allows is not carried out in the data analysis, I believe. In that case they should have modeled the effect of more components of the past $\bar{\mathbf{V}}(t)$. With such a nonparametric choice of censoring model, they argue that α is still not practically identified.

Suppose that the nonparametric model used in the sensitivity analysis models the dependence of $\Lambda_0(t|\bar{\mathbf{V}}(t))$ on $\bar{\mathbf{V}}(t)$ by extracting from the CD4 past several covariates. Suppose now that we fit a lower-dimensional nested model that simply sets the coefficients in front of several of these covariates equal to 0; for example, this model might include only CD4(t) as covariate. If this lower-dimensional model is correct, then the corresponding confidence interval for α gives a plausible range of α values for the more nonparametric model. Because there is no perfect solution, it makes sense to be satisfied with such a guessed plausible range of α values. To make the analysis more sophisticated, one could obtain such confidence intervals for α for a nested sequence of models, so that one also obtains an idea about how strong the confidence intervals of α depends on the actual assumed model for $\Lambda_0(t, \bar{\mathbf{V}}(t))$.

This suggests the following nonparametric sensitivity analysis procedure.

1. Select a lower-dimensional model for $\Lambda_0(t, \bar{\mathbf{V}}(t))$ nested in the actual used model for $\Lambda_0(t, \bar{\mathbf{V}}(t))$.
2. Estimate $(\alpha, \mu, \boldsymbol{\gamma})$ simultaneously. Use the .95% confidence interval for α as the plausible range of α values in the next step.
3. For the more nonparametric model for $\Lambda_0(t, \bar{\mathbf{V}}(t))$, carry out a sensitivity analysis as in the article.

Note that this data-adaptive approach of obtaining a plausible range of α values still allows using expert knowledge. Namely, if experts have strong knowledge on the censoring mechanism, one can use that knowledge to select a lower-dimensional model for $\Lambda_0(t|\bar{V}(t))$ in step 1.

ADDITIONAL REFERENCES

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997), "Coarsening at Random, Characterizations, Conjectures and Counter-Examples," in *Proceedings of the First Seattle Symposium in Biostatistics* 1995, eds. D. Y. Lin and T. R. Fleming, New York, Springer-Verlag, pp. 255–294.

Jacobsen, M., and Keiding, N. (1995), "Coarsening at Random in General Sample Spaces and Random Censoring in Continuous Time," *The Annals of Statistics*, 23, 774–786.

Robins, J. M. (1993), "Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers," *Proceedings of the Biopharmaceutical Section, American Statistical Association*, pp. 24–33.

Comment

Peter J. DIGGLE

I greatly admire the work embodied in this important article. The authors' mathematical skill provides a solution to a notoriously difficult methodological problem. Their statistical insight raises a challenge to parametric modelers who advocate other, less robust solutions. In the following remarks I take the general point of view that in comparing the semiparametric modeling approach of the present article with a parametric modeling approach, we are comparing shades of gray—both are modeling approaches, informed by a combination of scientific judgment and statistical formalism, differing only in the extent to which they choose to constrain the family of models under consideration and, by the same token, to admit modeling assumptions that go beyond the empirically verifiable.

In my experience, scientists are inveterate overoptimists. They ask questions of their data that their data cannot answer. Statisticians can either refuse to answer such questions, or they can explain what is needed over and above the data to yield an answer and be properly cautious in reporting their results. Parametric models represent a formal articulation of what I mean by "over and above the data." Sometimes the willingness to make untestable assumptions opens up new scientific insights; other times, it generates a misleading answer. To find out which of these two situations prevails might require independent confirmation by follow-up studies. This is an integral part of the scientific method (but perhaps sits uncomfortably alongside clinical trial ethics). A counterpart to the authors' "It is not what you do not know that hurts you; it is the things you think you know, but do not" is "We buy information with assumptions" (Coombs 1964).

What I like about the quotation from Coombs is its implicit invitation to the reader to consider the possibility that what is bought may be worth more *or* less than its price. Put another way, the extent to which information or assumptions over and above the data should be allowed to impact on statistical inference may depend on the type of question being asked. The more complicated the question (in conceptual terms), the more the balance may move in favor of parametric modeling. In the longitudinal setting of the article, if I want to estimate a marginal population mean (i.e., the average value that I would expect to see in a drop-out–free population), then I would prefer to avoid detailed assumptions about the longitudinal variation in the data, and the semiparametric approach is very attractive. If I want to look in more detail at longitudinal trajectories, either population-averaged or especially subject-specific (e.g., to make clinical decisions on individual patients), and adjusted for the effects of measurement error on the observed responses, then I think I am more or less forced to develop a parametric model, quite likely involving time-varying random effects to describe the individual subjects' trajectories. The data from the ACTG 175 trial, which the authors very kindly passed on to me, contain much more longitudinal information than the authors use to answer the simple but very relevant question: What would have been the mean response at 56 weeks in each treatment arm had there been no drop-outs? Had the authors sought to address more detailed questions about longitudinal trajectories, would their methodology have moved closer toward a parametric modeling approach?

At Lancaster, Rob Henderson and I are working on a modeling approach to problems of this kind that is still semiparametric but in a weaker sense than is used in the article. We postulate a parametric linear model for a longitudinal sequences of measurements Y_{ij}: $j = 1, \ldots, n_i; i = 1, \ldots, m$, where i indexes subjects, j indexes occasions within subjects, and t_{ij} denotes the time at which the measurement Y_{ij} is made. We assume that

$$Y_{ij} = x_1(t_{ij})'\beta + \mathbf{W}_{1i}(t_{ij}) + Z_{ij}, \tag{1}$$

where the $Z_{ij} \sim \mathrm{N}(0, \tau^2)$ are mutually independent measurement errors, the $\mathbf{W}_{1i}(\cdot)$ are independent copies of a nonstationary stochastic process, which is itself decomposable into a linear random-effects component and a serially correlated stationary component, and $x_1(t_{ij})$ is a possibly time-varying vector of covariates for Y_{ij}. For the drop-out model, or more generally for the intensity function of an associated point process of recurrent events, we assume an expression of the form

$$\lambda_i(t_{ij}) = \lambda_0(t_{ij}) \exp\{x_2(t_{ij})'\alpha + \mathbf{W}_{2i}(t_{ij})\}, \tag{2}$$

where $\lambda_0(\cdot)$ is a nonparametrically specified baseline intensity that is modulated by a log-linear regression including both fixed and random components through the terms $x_2(t_{ij})'\alpha$ and $\mathbf{W}_{2i}(t_{ij})$. In this model, association between the measurement and recurrent event histories of an individual subject is induced by dependence between the processes $\mathbf{W}_{1i}(\cdot)$ and $\mathbf{W}_{2i}(\cdot)$. There is an emerging literature

Peter J. Diggle is Director, Medical Statistical Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K. The author would like to thank Ray Carroll and all of the participants in the recent conference on Informative Missing Values held at Texas A&M University to mark Ray's 50th birthday. The two days of discussion at that conference helped him to clarify his thoughts on the issues raised by dropouts in longitudinal studies (although readers of these comments may feel, as he does that he still has some way to go). This research has been supported in part by National Institute of Mental Health grant R01 MH56639.

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

Table 1. Estimated Bias of $\hat{\mu}$ and Standard Error of the Estimated Bias, From a Simulation Experiment With n = 100 Subjects, and 1,000 Simulated Replicates, for Each of Several Values of τ

τ^2	.01	.1	.5	1.0	100.0
Bias	−.0036	−.0230	−.0594	−.0910	−.1735
SE	.0034	.0034	.0035	.0035	.0040

on semiparametric models of this general kind (for a recent review, see Hogan and Laird 1997).

Within the specific context of the problem that this article does address—estimating a marginal population mean—the core of the semiparametric method is, I think, the unnumbered equation in Section 3,

$$\hat{E}[l(Y)|\mathbf{V} = v] = \frac{1}{n_v}\sum_{i=1}^{n}\frac{\Delta_i I\ (\mathbf{V}_i = v)l(Y_i)}{\hat{\pi}(v, Y_i)} \qquad (3)$$

The intuitive interpretation of (3) is that an unweighted mean (which would give the required answer in a drop-out–free population) is modified by weighting each observation Y_i inversely according to the estimated probability $\pi(v, Y_i)$ that Y_i is observed conditionally on its covariate value v. The simulation studies in Section 5 of the article show that this can give good results, both in terms of unbiased estimation and accurate estimation of a standard error, in large samples. As the authors themselves point out, it could lead to difficulties if the $\hat{\pi}$ can get close to 0, and my guess is that these difficulties become relatively more acute in small samples.

I am especially intrigued by the authors' discussion in Section 7.2.3, concerning the strategy of including additional covariates in the drop-out model. Their conclusion, if I understand it correctly, that this does not help at all seems counterintuitive. I wonder, therefore, whether the failure of this strategy is a byproduct of the insistence on a fully nonparametric formulation of the dropout mechanism. A simple parametric counterexample would be to postulate a model in which $Y|U \sim \mathrm{N}(\mu + U, \sigma^2)$ and $U \sim \mathrm{N}(0, 1)$. The marginal mean of Y is μ. Now suppose that $\mathrm{logitP}(Y \text{ missing}|U) = \alpha + \beta U$. If U is unobserved, then this is an informative drop-out model; if U is observed (without error) as a covariate, then the model becomes a completely random drop-out covariate-dependent model (Little 1995), and the analysis would be straightforward. Perhaps more realistically, if in this model we observe not U itself, but a covariate X that is correlated with U, an analysis based on the (incorrect) assumption that drop-out is completely random but dependent on X; for example, that

$$\mathrm{logitP}(Y \text{ missing}|U) = \alpha + \beta X \qquad (4)$$

should alleviate to some extent the problem of dealing adequately with the informative drop-out. I have carried out a small simulation experiment using this model. The true values of the model parameters were $\mu = 0, \sigma = .1, \alpha = -.5$, and $\beta = .5$. Hence the drop-out mechanism is such that the probability of drop-out increases with Y, and an analysis that treats the drop-outs as ignorable is liable to give a negatively biased estimator for μ.

The observed covariate X was generated as $X = U + Z$, where $Z \sim \mathrm{N}(0, \tau^2)$; hence the correlation between U and X is $\rho = 1/\sqrt{1+\tau^2}$. Each experiment generates data from $n = 100$ subjects, and the experiment was replicated 1,000 times for each of several values of τ. Table 1 shows Monte Carlo estimates of the estimator $\hat{\mu} = n^{-1}\sum_{i=1}^{m} Y_i/\hat{\pi}_i$ where the estimates of $\pi_i = \mathrm{P}(Y_i \text{ observed})$ are obtained from the incorrect drop-out model (4).

When $\tau^2 \approx 0$ (so that $\rho \approx 1$), the observed X is a near-perfect surrogate for the unobserved U and the estimator for μ is unbiased. As τ^2 increases, negative bias develops as predicted, but this development is progressive. Provided that τ^2 is small (i.e., a good surrogate can be found for the random effect U), the strictly incorrect analysis assuming completely random, covariate-dependent drop-out can give approximately correct inferences. At the opposite extreme, when τ^2 is large, the observed covariate X is unrelated to the drop-out mechanism, and adjustment for it yields no benefit.

Whatever modeling strategy is adopted, I think there would be general agreement on all of the following:

- When informative drop-out cannot be ruled out, sensitivity analyses are preferable to placing total faith in a single fitted model.
- The question of plausible ranges for sensitivity analysis parameters is both important and difficult.
- The ideal outcome, that substantive inferences are robust to variation of sensitivity parameters over the whole of their permissible ranges, may not be achieveable in practice.

These considerations lead me to conclude that sensitivity parameters should, if possible, have an interpretation which is readily explainable to the scientist whose data are being analyzed. This raises a (possibly tenuous), analogy with the elicitation of prior distributions for Bayesian inference. However, although it may be true that in practice, proponents of Bayesian inference are also naturally attracted to parametric modeling formulations, my own view is that the question of whether parametric modeling based on assumptions "over and above the data" is a good *modeling* strategy is quite separate from the issues that distinguish Bayesian from non-Bayesian *inference.*

ADDITIONAL REFERENCE

Coombs, C. H. (1964), *A Theory of Data*, New York: Wiley.

Comment

Roderick J. LITTLE and Donald B. RUBIN

1. INTRODUCTION

We appreciate the opportunity to comment on the article by Scharfstein, Rotnitzky, and Robins (henceforth SRR), which is an ambitious and thought-provoking attempt to solve a difficult methodological problem with mathematical dexterity. In this comment we consider four aspects of SRR's work: the role of sensitivity analysis, the importance of using all available data on drop-outs to reduce the impact of nonignorable nonresponse, the form of the SRR model, and methods of estimation and inference.

2. SENSITIVITY ANALYSES FOR NONIGNORABLE NONRESPONSE

Nonignorable missing data (Rubin 1976) pose a difficult problem, because the data do not provide information about parameters characterizing nonignorable aspects of the missing-data mechanism, at least without making assumptions untestable from the data at hand. Early work, particularly in econometrics, attempted to estimate simultaneously parameters of the complete-data model and parameters characterizing nonignorable nonresponse (Amemiya 1984; Heckman 1976; Nelson 1977); a more recent application of this approach in the repeated-measures setting is that of Diggle and Kenward (1994). Many authors, however, have criticized this approach (Copas and Li 1997; Glynn, Laird, and Rubin 1993; Little 1985, 1994a; Rubin 1994; Tukey 1986), because the models are identified purely on the basis of normal distributional assumptions, or assumptions that particular regression coefficients are exactly 0. The estimates from these models are highly sensitive to minor deviations from assumptions, such as lack of normality or a particular regression coefficient being close to 0 rather than 0. For example, the application of a nonignorable selection model to income nonresponse in the Current Population Survey (Lillard, Smith, and Welch 1986) yielded predictions for the nonrespondents that differed drastically from estimates based on independent data sources (David, Little, Samuhel, and Triest 1986). We believe that statisticians and economists have generally moved away from these approaches.

We agree with SRR that a sensitivity analysis is a rational approach to nonignorable nonresponse, and have advocated this approach in our own work (Little 1994b; Little and Rubin 1987; Little and Wang 1996; Rubin 1977). When sufficiently transparent to be understandable by substantive researchers, sensitivity analysis has also been used by these researchers to help elucidate possible consequences of nonignorable missing-data mechanisms; see, for example, the application by Connors et al. (1996) of the sensitivity analysis methods of Rosenbaum and Rubin (1985).

Nevertheless, sensitivity analysis is not without problems. First, in many practical settings users of statistics greatly favor simplicity and concision in the presentation of results. It is often hard enough to convince them of the need to go beyond point estimates and present confidence intervals for estimands of interest, much less the need for presenting a range of estimates with two different types of uncertainty, sampling error within a particular model and variability of estimates across models. Second, a sensitivity analysis usually needs to be confined to a relatively small number of parameters (say one or two), as otherwise the set of answers obtained by simultaneously varying a set of parameters may become overwhelming. Third, many different forms of sensitivity analysis can be contemplated and may yield contradictory conclusions.

3. LIMITING THE SCOPE OF NONIGNORABLE MODELING

Given the problems inherent in nonignorable modeling, we have generally advocated trying to make the ignorability assumption as plausible as possible by collecting as much information about incomplete cases as possible, and then including this information for inferences via model-based analyses, such as by multiple imputation. In fact, we believe that in situations where good covariate information is available and included in the analysis, the missing at random (MAR) assumption may often be a reasonable approximation to reality, thus obviating the need for a sensitivity analysis to model nonignorable nonresponse. For example, both David et al. (1986) and Rubin, Stern, and Vehovar (1996) presented examples where the straightforward MAR model predicts actual outcomes better than standard and arguably plausible nonignorable models.

In the repeated-measures setting studied by SRR, a useful positive feature for MAR modeling is the availability of repeated measures on subjects prior to drop-out, which can be used together with other covariate information to generate a predictive distribution for the missing values under the MAR assumption. Thus our philosophy to modeling the data is to make full use of this information to reduce the partial association between drop-out and the outcome variables of interest. In cases where the information measured for drop-outs is judged insufficient to account for differences between those cases and the cases that remain in the

Roderick J. Little is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48105 (E-mail: *rlittle@umich.edu*). Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *rubin@hustat.harvard.edu*).

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

study, the MAR analysis can be supplemented by clearly formulated sensitivity analyses based on scientifically plausible nonignorable models.

It is here that we part company with SRR, who in their attempt to minimize parametric assumptions effectively make limited use of covariate information. This results in increased sensitivity of inference to the nonignorable component of the model, and possibly overly conservative inferences. We agree in principle with SRR that there is some interest in seeing how much uncertainty is engendered by an analysis that makes minimal assumptions; however, assumptions are inevitable when handling missing data in practice. Bounds based on a worst-case analysis (e.g., Cochran 1963, sec. 13.2; Horowitz and Manski 1998) are usually too wide to be useful, except when the amount of missing data is trivially small. Coding the repeated measures prior to drop-out as categorical covariates in SRR's model of (1) and including all interactions drastically limits the number that can be accommodated. Hence the more constrained SRR model of (2) seems much more practical in repeated-measures settings. Even the model of (1) assumes that there are no interactions between the effects of the final outcome and the covariates on the probabilities of drop-out.

4. SCIENTIFIC RATIONALE FOR THE ADOPTED MODEL

When nonignorable models are applied, it is critical to link the model for the missing-data mechanism to whatever science is known about the problem under study. Consequently, we feel that it is imperative that nonignorable modeling be as transparent as possible, in the sense that the underlying assumptions can be readily appreciated by users who understand the data but may not be professional statisticians. To help understand the meaning of the SRR model and link it to previously proposed models, we consider their basic equation (1) for the special of monotone missing data with $T = 3$ discrete evenly spaced times of measurement, no between-subject covariates, and categorical repeated measures. In that case, the full model can be specified in terms of the joint distribution of Q, Y_1, Y_2 and $Y_3 \equiv Y$, where $Q = j$ if a subject drops out between measurement times j and $j + 1$ and Y_j is the outcome at time j. The discrete analog of the model (1) has the form

$$1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = (1 - \text{pr}(Q = 1|Y_1))\Psi(\alpha_0 Y_3),$$

$$1 - \text{pr}(Q = 2|Y_1, Y_2, Y_3, Q > 1) = (1 - \text{pr}(Q = 2|Y_1, Y_2, Q > 1))\Psi(\alpha_0 Y_3),$$

and

$$\text{pr}(Q = 3|Y_1, Y_2, Y_3, Q > 1) = 1 - \text{pr}(Q = 2|Y_1, Y_2, Y_3, Q > 1),$$

where $\Psi(u) = \exp(\exp(u))$ corresponds to a complementary log-log link function. This model implies that the probability of dropping out at time 2 ($Q = 1$) depends on the outcomes at times 1 and 3 but not on the outcome at time 2, and the probability of dropping out at time 3 ($Q = 2$) depends on the outcomes at times 1, 2, and 3. Furthermore, the parameter α_0 governing the dependence of drop-out on the outcome at time 3 is constrained to have the same value for both drop-out times. This formulation does not appear to be a very plausible missing-data mechanism for any data that we can think of. Moreover, because α_0 is a parameter whose meaning changes with every drop-out pattern, it is very difficult to interpret sensitivity analyses with different values of α_0.

A more plausible form of the model might be

$$1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = (1 - \text{pr}(Q = 1|Y_1))\Psi(\alpha_0 Y_2)$$

and

$$1 - \text{pr}(Q = 2|Y_1, Y_2, Y_3, Q > 1) = (1 - \text{pr}(Q = 2|Y_2, Q > 1))\Psi(\alpha_0 Y_3),$$

where dropping out at a particular time point depends on the value of the outcome at the time of drop-out and at the previous time. A natural parametric form of this model is

$$1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = \Psi(\gamma_1 + \alpha_0 Y_2 + \alpha_1 Y_1)$$

and

$$1 - \text{pr}(Q = 2|Y_1, Y_2, Y_3, Q > 1) = \Psi(\gamma_2 + \alpha_0 Y_3 + \alpha_1 Y_2),$$

which is closely related to the drop-out model of Diggle and Kenward (1994), although differing in the choice of the complementary log-log rather than the logistic link. Unlike the estimation approach of Diggle and Kenward, we would advocate a sensitivity analysis based on this model with prespecified values of α_0.

Given the large class of nonignorable models, we believe that they need to be assessed in the context of particular applications, and we have some comments about the specific application of the SRR model to the AIDS clinical trial data. In particular, the analysis treats death and noncompliance to treatment in the same way as other forms of drop-out; for example, effectively imputing CD4 counts to subjects after they have died. The scientific rationale for this approach seems questionable. A more reasonable approach to noncompliance is to distinguish subjects by underlying compliance type under both treatment arms (Angrist, Imbens, and Rubin 1996); to deal with censoring due to death, the key piece of information is the underlying true survival type under both treatment arms (Rubin 1998, sec. 6). These questions can both be usefully formulated as missing-data problems, but they require different models for missing data than the model for missing CD4 counts of those who survived and complied with treatment.

5. INFERENCE PROCEDURES UNDER THE CHOSEN MODEL

Our work has typically adopted a standard likelihood-based approach to inference, whereas SRR base estimation on generalized estimating equations, whose basic form is given in their (3). We first discuss this estimation approach for the special case where the function b equals 0.

Remark 1. If the data are missing completely at random (MCAR), then (3) reduces to the complete-case mean and hence discards the incomplete cases. Under MCAR, this estimator is unbiased but involves a loss of efficiency, because the data in incomplete cases are ignored. In repeated-measures settings the loss of information can be quite significant, especially when the data on the covariates and outcome data prior to drop-out are highly predictive of Y.

Remark 2. If the data are MAR, then (3) reduces to the weighted mean of the complete cases, where the weights are estimates of the inverse probabilities of drop-out given the covariates and outcomes prior to drop-out. This estimator is commonly applied to unit nonresponse in surveys (Little and Rubin 1987, sec. 4.4) and is based on the ideas underlying the Horvitz–Thompson estimate (Horvitz and Thompson 1952). The method is also a weighting analog of the multiple imputation method of Lavori, Dawson, and Shera (1995), which is implemented in the first release of SOLAS (Statistical Solutions, Inc. 1998). The incomplete cases are used for bias adjustment but, as in the MCAR case, are not used to predict missing values of Y. Thus, whereas the SRR estimator is focused on bias reduction, model-based approaches have the potential to reduce both bias and variance, although they can be vulnerable to misspecification of the regression of the missing variables on the observed outcomes and covariates.

Remark 3. If the data are not MAR, then the weights in (3) are modified to yield consistent estimates under the assumptions of (1). As in the MAR case, the information in the incomplete cases is confined to bias adjustment, and we suggest that (3) is inefficient when the outcomes observed prior to drop-out are good predictors of the missing outcomes, appeals to semiparametric efficiency bounds notwithstanding.

Perhaps the potential inefficiency of the estimator (3) with $b = 0$ can be alleviated by the inclusion of a nonzero function b, which allows inclusion of the incomplete data in the estimating equation. However, we have considerable difficulty following the prescriptions described in the article for choosing b, and the assumptions implicit in the choice. We suggest that more work is needed to clarify the choice of b, even in simple cases of the model.

ADDITIONAL REFERENCES

Amemiya, T. (1984), "Tobit Models: a Survey," *Journal of Econometrics*, 24, 3–61.

Angrist, J., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–472.

Cochran, W. G. (1963), *Sampling Techniques* (2nd ed.), New York: Wiley.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. (1996), "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association*, 276, 889–897.

Copas, J. B., and Li, H. G. (1997), "Inference for Non-Random Samples" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 55–97.

David, M. H., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1986), "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association*, 81, 29–41.

Glynn, R., Laird, N. M., and Rubin, D. B. (1993), "Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-Ups," *Journal of the American Statistical Association*, 88, 984–993.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Horowitz, J. L., and Manski, C. F. (1998), "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37–58.

Lavori, P. W., Dawson, R., and Shera, D. (1995), "A Multiple Imputation Strategy for Clinical Trials With Truncation of Patient Data," *Statistics in Medicine*, 14, 1913–1925.

Lillard, L., Smith, J. P., and Welch, F. (1986), "What do We Really Know About Wages: The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, 94, 489–506.

Little, R. J. A. (1994a), Discussion of "Informative Drop-Out in Longitudinal Data Analysis" by P. Diggle and M. G. Kenward, *Applied Statistics*, 43, 85–85.

——— (1994b), "A Class of Pattern-Mixture Models for Normal Missing Data," *Biometrika*, 81, 471–483.

Little, R. J. A., and Wang, Y.-X. (1996), "Pattern-Mixture Models for Multivariate Incomplete Data With Covariates," *Biometrics*, 52, 98–111.

Nelson, F. D. (1977), "Censored Regression Models With Unobserved Stochastic Censoring Thresholds," *Journal of Econometrics*, 6, 581–592.

Rosenbaum, P. R., and Rubin, D. B. (1985), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1994), Discussion of "Informative Drop-Out in Longitudinal Data Analysis" by P. Diggle and M. G. Kenward, *Applied Statistics*, 43, 80–81.

——— (1998), "More Powerful Randomization-Based p Values in Double-Blind Trials With Noncompliance," *Statistics in Medicine*, 17, 371–385.

Rubin, D. B., Stern, H., and Vehovar, V. (1995), "Handling 'Don't Know' Survey Responses: The Case of the Slovenian Plebiscite," *Journal of the American Statistical Association*, 90, 822–828.

Statistical Solutions, Inc. (1998), *SOLAS Program for Missing Data Analysis*, Cork, Ireland: Author.

Tukey, J. W. (1986), Comment on "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes" by J. J. Heckman and R. Robb, in *Drawing Inferences from Self-Selected Samples*, ed. H. Wainer, New York: Springer-Verlag.

Comment

Nan M. LAIRD and Donna K. PAULER

1. SELECTION MODELS

Selection models offer an intuitive approach to dealing with the difficult problem of nonignorable nonresponse. It has long been understood that both the data model and the nonresponse model are not completely identifiable from data, but there has been much difficulty in determining which particular models can be be estimated from data, and, for an estimable model, which aspects of the model are well estimated from data and which aspects are sensitive to the model assumptions.

The path taken by the authors is to be completely nonparametric in estimating the mean response, assuming nothing about the parametric distribution of outcome or how it depends on the other variables, observed or unobserved; they assume a semiparametric form for the nonresponse model, where the part depending on the observables is allowed, where feasible, to be unspecified and then posit a model with known parameters for the dependence of nonresponse on the unobserved variables. The attractive feature of their approach, at least in relatively simple settings, is that one fully utilizes the observed data to the maximum extent possible, and the nonresponse parameter (and the assumed model for nonresponse) can be varied to study sensitivity to assumptions. In most real problems, it will not be possible to fully specify the nonresponse model as a function of observables, and the authors suggest some simplification of the nonresponse model. However, they continue to recommend to specify the parameter determining the dependence of the nonresponse model on the unobserved outcome, even though there is now some information in the data about this parameter.

Several aspects of this approach are readily understood by considering the very simple setting where all of the variables are categorical, only one variable is subject to nonresponse, and there is only one time of nonresponse. For this setting, it is straightforward to see that fitting a saturated model for the data and a nonresponse model that is saturated in the observed variables leaves 0 degrees of freedom in the data for estimating dependence of nonresponse on the outcome subject to missingness (Baker and Laird 1988). Still leaving the data model saturated, but putting a structure on the nonresponse model that is not fully saturated in the observed covariates, leaves positive degrees of freedom, which permits estimation of all of the model parameters, although in practice it can difficult to determine which models are estimable, and the likelihoods may be quite flat. This is analogous to the authors model $B(\alpha_0)$, and we can appreciate the point of view that one should continue to specify and not estimate the parameter α_0. But in the more realistic case where the outcome interacts with covariates in the nonresponse model, some strategy that combines sensitivity analysis with estimation may be desired. To this end, with categorical data at least, likelihoods are a very useful way of exploring goodness of fit and model sensitivity, and we feel that the present approach could benefit with some type of objective function that could be used for this purpose.

Another feature of the proposed method also arises with maximum likelihood (ML) estimation and categorical variables. Baker and Laird (1988) showed that when estimating nonignorable nonresponse models for the $2\times2\times2$ table, the ML solution will sometime lie on a "boundary" in the sense that all values of the outcome for the nonrespondents will be "imputed" to be either 0 or 1. A similar phenomenon was observed in larger tables. Here, even though the model may be saturated in the sense that the number of parameters equals the number of degrees of freedom, there is not a perfect fit to the data, in the sense that $(O - E) \neq 0$ for those margins that are observed.

Although the authors are very careful to note throughout that the resulting estimators depend not only on the sensitivity parameter, but also on the nonresponse model, certain features of their analysis of the CD4 data invite the reader to feel, as indeed the authors tell us to feel, that "this conclusion is quite robust. Significant differential selection biases would have to occur to alter this conclusion" (that AZT + ddi is to be preferred over AZT). The point is, of course, that within the context of the given model, the conclusion is robust. One particular feature of their methods that invites this feeling of complacency about the conclusions is the implicit assumption, discussed at length by the authors, that the outcome for a nonresponder is bounded by the outcomes of the responders. In many situations this may not be a reasonable assumption. The other obvious feature of the analysis that is not discussed is that the parameter α_0 may be a vector, and the nonresponse model may depend not only on the outcome, but also on its interaction with observed covariates. Particularly in this example, it would seem desirable to interact time of dropout with outcome in the nonresponse model.

2. PATTERN-MIXTURE MODELS

An alternative factorization of the model for complete data is the pattern-mixture model, which is briefly introduced in Section 7.3.2 but perhaps deserving of more at-

Nan M. Laird is Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Donna K. Pauler is Assistant Professor, Biostatistics Center, Massachusetts General Hospital, Boston, MA 02114.

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

tention. Denoting the complete data for a subject by $C = (Y, \mathbf{V}_Q^-, \mathbf{V}_Q^+, Q)$, where $\mathbf{V}_Q^+$ denotes time-varying covariates occurring after drop-out Q, $\mathbf{V}_Q^-$ denotes those occurring before drop-out, and Y denotes the primary endpoint of interest occurring at the end of the study, the pattern-mixture model can be defined as

$$F_C = F_{Y|\mathbf{V}_Q^+,\mathbf{V}_Q^-,Q} F_{\mathbf{V}_Q^+|\mathbf{V}_Q^-,Q} F_{\mathbf{V}_Q^-|Q} F_Q. \quad (1)$$

From this factorization, it is easy to see that F_Q, all remaining three pieces for completers, and the margins $F_{\mathbf{V}_Q^-|Q}$ for noncompleters are identifiable. The data contain no information about the remaining portions of the model.

Various likelihood-based approaches for drawing inference from pattern-mixture models have been suggested. Rubin (1977) discussed Bayesian techniques for utilizing subjective information to relate effects of nonrespondents in sample surveys to those of respondents. Little (1993) obtained identifying restrictions in simple bivariate normal samples by specifying restrictions corresponding to the presumed operational selection model. Hogan and Laird (1997) handled the lack of identifiability by making specific assumptions about the relationship of outcome in drop-outs and completers.

In the likelihood case, an advantage of pattern-mixture models is that they are not as sensitive to distributional assumptions as selection models, where in the latter, estimates of parameters for the complete data may not be robust to misspecification of the selection mechanism or model for the unobservables (Brown 1990; Glynn, Laird, and Rubin 1993; Little 1982, 1985), and estimates of the parameters of the selection mechanism may be driven almost completely by the assumed complete-data distribution (Little and Rubin 1987, chap. 11). From the Baker and Laird (1998) model, it is clear that results in the selection model mechanism can be equally driven by the assumed model for nonresponse. In contrast, the model for the nonresponse in mixture models can be estimated completely from observed data, as can the model for the complete data, conditional on being a completer. Although (1) is expressed it terms of distributions, if only the mean of Y is of interest, then the modeling assumptions will be needed only for $E(Y|\mathbf{V}_Q^-, Q)$. In some settings it may be more natural to specify a model for this conditional expectation rather than a model for the nonresponse in terms of the outcome Y. As we illustrate in the next section, the assumptions necessary to implement the mixture model are considerably more transparent than those needed for the selection model, and we find the results easier to explain and interpret.

3. SIMPLE SEMIPARAMETRIC PATTERN-MIXTURE MODELS

Based on expression (1), it is easy to derive crude but simple nonparametric estimators for the identifiable portions of the model and to insert easily interpretable assumptions about the nonidentifiable parts. Without much loss of information, F_Q may be estimated by grouping drop-out times into bins, and $F_{\mathbf{V}_Q^-|Q}$ may be estimated using the empirical distribution function stratified across the bins. As in the authors' estimator, for Q and $\mathbf{V}$ of suitably low dimension, the approach can be completely nonparametric for all identifiable portions of the model. We outline two example to illustrate how linear estimators of the mean may be formed.

Example 1. Suppose that $T = 2$ and drop-outs can occur only at T, so that $Q = 1$ denotes those who miss measurement 2 and $Q = 2$ denotes completers. A covariate $\mathbf{V}_1$ is collected at time 1 and the primary endpoint Y is collected at time 2. Assume a linear model for the mean of Y for completers, $E(Y|\mathbf{V}_1, Q = 2) = \beta_0 + \beta_1 \mathbf{V}_1$ and that the mean of those who drop out at time 1 differs by the constant ϕ from the completers: $E(Y|\mathbf{V}_1, Q = 1) = E(Y|\mathbf{V}_1, Q = 2) + \phi$. Then, by averaging over the appropriate empirical distributions, the marginal mean of Y is calculated as $\mu_0 = E(Y) = \beta_0 + \beta_1 \bar{\mathbf{V}}_1^{\mathrm{AC}} + \pi\phi$, where the superscript AC denotes the mean over all available cases and π is the proportion of subjects with $Q = 1$. An estimate of μ_0 is obtained by substituting estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ from the complete cases and the sample proportion $\hat{\pi}$ of drop-outs. Because $\hat{\mu}_0$ is a function of $(\hat{\beta}_0, \hat{\beta}_1, \bar{\mathbf{V}}_1^{\mathrm{AC}}, \hat{\pi})$, its variance may be calculated from the variance of these estimates using the delta rule.

Example 2. One may generalize the model in Example 1 for the data from ACTG175 by discretizing the drop-out times at weeks 8, 20, 32, 44, and 56 (completers) and assuming a linear dependence of Y on covariates for completers. There are many ways to relate the conditional means of noncompleters, $E(Y|Q = t)$, to that of completers, $E(Y|Q = T)$. We suggest using the authors' relationship (16), which compares the means of those who drop out at time t to those who continue at time t. Alternatively, one may compare to those who drop out at time $t + 1$, or to the completers directly. In the latter case, it is sensible to include an interaction with time, because for a given set of covariates, the means of those who drop out later in the study should be more similar to the means of completers than those who drop out earlier.

4. SUMMARY

If selection bias is suspect, then the statistician must collaborate with the investigator to form subjective notions about the nature of the possible selection mechanism. This article makes a valuable contribution by explicating which parts of the operational selection model or pattern-mixture model are estimable from the data and which are not, preventing subjective opinion from imposing hidden biases. However, being equipped with tools to determine sensitivity to selection bias does not free the investigator from the need to try to design against nonignorable dropout in new studies. Indeed, perhaps the real value in formal hypothetical models for dropout lies in their ability to inform practitioners of the dangers of selection bias.

ADDITIONAL REFERENCES

Baker, S., and Laird, N. M. (1988), "Regression Analysis With Categorical Data Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association*, 88, 62–69.

Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies," *Biometrics*, 46, 143–157.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993), "Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups," *Journal of the American Statistical Association*, 88, 984–993.

Little, R. J. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, 77, 237–250.

——— (1994), "A Class of Pattern-Mixture Models for Normal Incomplete Data," *Biometrika*, 81, 471–483.

Rubin, D. B. (1974), "Characterizing the Estimation of Parameters in Incomplete-Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

Rejoinder

Daniel O. SCHARFSTEIN, Andrea ROTNITZKY, and James M. ROBINS

1. INTRODUCTION

We thank the editor for organizing this discussion and the discussants for their stimulating comments. In their discussion, Fan and Zhang (FZ) elegantly point out the daunting uncertainties that exist when there is substantial drop-out and suggest that possibly "one should not adjust at all for drop-out bias if no reliable method is available for modeling the drop-out time." Although sympathetic with this viewpoint, we would not wish to discard costly and potentially important data without first taking a careful look. So what, if anything, can be done? For failure time outcomes or outcomes with a bounded range (e.g., dichotomous outcomes), the comparison of worst-case bounds is an obvious first step. If these are too wide to be useful, then a nearly nonparametric sensitivity analysis can help the investigators examine the stability of their conclusions under varying assumptions. As David Freedman concludes, "when substantial amounts of data are missing, the only analysis that matters is often a sensitivity analysis." Because the nonignorable selection bias function is at best only weakly identified, subjective input from subject matter experts is needed. Van der Laan fears that the task we have set for these experts is undoable and proposes a different, but related approach. Diggle endorses our approach for estimating simple functionals such as the mean, but suggests a more parametric approach when estimating complex functionals.

Laird and Pauler (LP) suggest an alternative approach based on fitting regressions to several "pattern-mixture" model variants. Little and Rubin (LR) agree that in some cases a sensitivity analysis is important, but propose fitting fully parametric models. They argue that the greater efficiency outweighs the associated potential for bias. LR even suggest that the sampling variability from a single parametric model might often suffice as a measure of uncertainty. We do not agree with LR's view that the user's desire for "simplicity and concision" helps justify such a limited inference. Our goal is to help scientists interpret the evidence in their data. By ignoring model uncertainty, we do a disservice to them, to the statistical profession, and to the science. If Diggle is correct in his opinion (which we share) that scientists are "inveterate overoptimists," then sensitivity analysis may serve as an important corrective.

If the only reason for censoring is loss to follow-up (rather than death or departure from the treatment protocol), then there is a reliable (albeit expensive) way to validly correct for selection bias. Specifically, just after the time at which the outcome of interest is to be measured, an extensive effort can be made to find and measure the outcome on a random sample of the drop-outs. We discuss this approach further in Section 3.4.

2. RESPONSE TO DISCUSSANTS

2.1 Freedman

Freedman considers the discrete time, single-occasion version of our nonparametric selection model (1). He provides an elegant, rigorous derivation and explicit characterization of the map that takes the law of the observed data and the nonidentified selection bias parameter to the unique law of the full data. In Section 5 of a related work, Robins et al. (1999), and Appendix A of our article, we provide less elegant derivations of this map in the discrete time multiple-occasion and continuous-time versions of the model.

2.2 Fan and Zhang

FZ are concerned that our selection model may be restrictive because it assumes a proportional hazards model for dropout. But this concern is unnecessary, because the general form (13) of our model includes all nonresponse mechanisms for some choice of the selection function r in (13). However, FZ make the further point, with which we agree and address further in Section 2.5.4, that it is difficult to choose among the possible selection functions, because the data offer no guidance; the selection function represents selection bias due to unmeasured factors and thus is not identified, unless we were to impose further, possibly incorrect, a priori restrictions such as (2). FZ hone in on this key point when they inquire about the possibility of

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

using the data to test the fit of our model (1), which is equivalent to testing whether the true selection function has the (log) linear form $\alpha_0 Y$. As we discuss in Section 2.3 of the article, whatever the true selection function, the (log) linear model (1) always fits the data perfectly and cannot be rejected by any statistical test. This is nonidentifiability. Because of this nonidentifiability, the observed data cannot help determine the magnitude or form of selection bias due to unmeasured factors without additional knowledge.

2.3 van der Laan

To circumvent the need for expert input, van der Laan proposes temporarily using a lower-dimensional semiparametric model to estimate a "plausible" range for the selection bias parameter α_0. Although this approach is ingenious, we have reservations about it. If the lower-dimensional model restrictions are not too strong, then α_0 will be weakly identified; as a result, in moderate-sized samples the resulting confidence interval for α_0 will be too wide to provide any useful guidance. Furthermore, in our experience it will be difficult to find a solution to the joint set of estimating equations. With stronger model restrictions, a solution can be found, but the resulting confidence interval for α_0 may be centered inappropriately and be misleadingly narrow because of incorrect functional form restrictions. In summary, although sympathetic with van der Laan's desire to circumvent expert input, we do not believe it possible.

2.4 Diggle

2.4.1 Complex Questions. We agree with Diggle's comments. In particular, we agree that the use of nonparametric or near-nonparametric sensitivity analysis should not preclude reporting the results of additional analyses based on more restrictive models, as long as one comments on the consistency of the results and on the strength and weakness of each approach. As discussed further in Section 3.2.8, our approach easily generalizes to include arbitrarily complex models for the law of the complete data, including random effects and semiparametric models. However, in our approach it is necessary to encode nonignorability by modeling the nonresponse probabilities (i.e., the hazard of censoring) as a function of the data (such as the outcome Y), rather than as a function of never-observed random effects.

2.4.2 Usefulness of Additional Covariates. Although we used the same symbol, the meaning of the sensitivity parameter α_0 in (1) changes as we change the covariates $\bar{\mathbf{V}}(t)$. Our Section 7.2.3 concerns itself with how to map a plausible range for the selection bias parameter of a model with many covariates to a range for the parameter of a model with fewer covariates. Due to somewhat poor exposition on our part, Diggle misinterpreted us as having concluded that the additional covariates do not help in estimating the mean μ_0 of Y. In fact, the availability of additional data will generally serve to decrease uncertainty and narrow bounds. Figures 1 and 3 serve as empirical examples. We use the asymptotes of the curves to approximate the bounds. When we have data only on IV drug user status, the approximate bounds in Figure 1 are $280 \leq \mu_0 \leq 510$ in the ddI arm; with the inclusion of additional covariates in Figure 3, the bounds narrow to $300 \leq \mu_0 \leq 470$. Similar narrowing occurs for other treatments.

2.5 Little and Rubin

2.5.1 Bounds. LR argue against reporting bounds based on worst-case analyses because "except when the amount of missing data is trivially small, these bounds are usually too wide to be useful." But, we view this problem as the reason for reporting bounds (in conjunction with other analyses): Wide bounds make clear the degree to which health decisions are dependent on combining the data evidence with prior beliefs. Indeed, by varying α_0, analyses based on either a nonparametric selection model or a near-nonparametric selection model have the potential to merge the worst-case bounds analysis ($\alpha_0 = \infty$ or $\alpha_0 = -\infty$) with nonignorable estimates ($\alpha_0 \neq 0$ but finite) and ignorable estimates ($\alpha_0 = 0$), while adjusting for measured covariates $\mathbf{V}(t)$ that explain drop-out.

2.5.2 Inference Procedures Under the Chosen Model. In their section 3 LR argue that in an attempt to minimize parametric assumptions, we are not making effective use of covariate information. This is incorrect, because locally efficient semiparametric estimators optimally combine the information contained in all of the data (covariates included) with the a priori restrictions on the distribution of the data encoded in the semiparametric model. The only way to obtain greater efficiency is to assume that the distribution of the complete data follows a parametric model. The problem is that imposing a parametric model results in bias if (as will essentially always be the case) the parametric model for the joint distribution of the measured covariates and the outcome Y is misspecified.

LR demonstrate that in the model characterized by (1) and (2), the augmented inverse probability of censoring weighted (AIPCW) estimator $\hat{\mu}(\hat{b})$ with augmentation function $\hat{b}$ identically 0 is similar to the Horvitz–Thompson estimator and, like the Horvitz–Thompson estimator, can be quite inefficient. They then appear to suggest that all AIPCW estimators will be quite inefficient, "appeals to semiparametric efficiency notwithstanding." But this argument is specious, because, as LR acknowledge in the final paragraph of their section 4, a locally efficient semiparametric estimator is an AIPCW estimator with a nonzero optimal augmentation function. Indeed, Robins et al. (1994) previously noted the relationship of the unaugmented IPCW estimator to the Horvitz–Thompson estimator and introduced augmented IPCW estimators specifically for the purpose of improving efficiency. As one example, Robins and Wang (1998) reanalyzed a dataset with missingness by design provided by Clayton, Spiegelhalter, Dunn, and Pickles (1998) and found that the estimated relative efficiency of a locally efficient semiparametric AIPCW estimator compared to a fully parametric maximum likelihood estimator (MLE) was .85, whereas that of the estimator with augmentation function identically 0 was only .16. Theoretical efficiency calculations show that, in conflict with the speculations of LR,

such results often can be expected when the observed covariates are highly correlated with missing outcomes. Perhaps the ultimate proof that AIPCW estimators can be efficient is that, as we show in Section 3.2.7 here, the MLE in a fully parametric model is itself an AIPCW estimator.

LR state in their section 4 that parametric "model-based approaches have the potential to reduce both bias and variance." We disagree. Variance reduction is offset by an increased potential for bias due to model misspecification. Specifically, in Section 3.2 here we show that in nonignorable models, whenever either the parametric MLE or Rubin's parametric multiple imputation estimator are consistent, then all AIPCW estimators are also consistent, but the converse is false. Furthermore, we show that in ignorable models the preceding statement is true for locally efficient AIPCW estimators.

Interestingly, in ignorable models, the above statement need not hold for inefficient AIPCW estimators. Thus, somewhat surprisingly, in ignorable models locally efficient AIPCW estimators prevent bias as well as increase efficiency relative to other AIPCW estimators. The adaptive AIPCW estimator proposed in Section 4.3 and Appendix C of our paper is locally efficient under ignorability.

2.5.3 Limiting the Scope of Nonignorable Models. LR state that when good covariate information is available and included in the analysis, "the MAR assumption may often be a reasonable approximation to reality, thus obviating the need for a sensitivity analysis to model nonignorable nonresponse." We disagree. It seldom would be possible to determine when the missing at random (MAR) assumption is a reasonable approximation, as there may be important unmeasured common causes of drop-out and the outcome. Hence reasonable people may disagree as to whether the sensitivity parameter α_0 is close to its MAR value of 0. In fact, the logic of LR's argument is unchanged if we replace LR's quoted words with the words "ignorability of treatment assignment conditional on covariates may often be a reasonable approximation to reality, obviating the need for a randomized trial." Many epidemiologists have made this argument with regard to the apparent protective effect of beta carotene on lung cancer seen in many observational studies. But recent randomized trials have shown that beta carotene causes, rather than protects against, lung cancer. Of course, a sensitivity analysis, unlike a randomized experiment, cannot provide a definitive answer; rather, it is but a sober warning that the true uncertainty may be much greater than the sampling variability associated with a single ignorable model. See Section 3.5.1. for further discussion of these issues.

2.5.4 Scientific Rationale for the Adopted Model. Choice of Functional Form. LR and LP criticize our choice in (1) of the linear form $\alpha_0 Y$ for the selection function. They say our choice assumes (a) that drop-out at t only depends on the possibly unobserved future through the final observation Y, (b) no interactions of Y with other covariates are included, and (c) the selection parameter α_0 is a scalar. However, as we explained near the end of our Section 7.2.2, we chose $\alpha_0 Y$ to illustrate the proposed methodology not because "we thought it substantively plausible, but rather because it is the usual default choice ..." More substantively motivated choices would be necessary for a complete analysis of the data.

Implausibility of Little and Rubin's "Plausible" Model. In Sec. 7.3.1 of our article, we criticized our model (1) as implausible because it assumes that drop-out depends on the entire future only through the final observation. We thus recommended the alternative model (15), which we called model $A^*(\alpha_0)$. Model (15) overcomes the deficiencies of (1) by modeling the conditional hazard of drop-out given only the observed past covariates and the final observation Y, rather than the entire future. In their section 3, LR repeat our criticism of model (1), having apparently overlooked our discussion in Section 7.3.1. Then, rather than adopting our recommended model (15), they suggest an alternative model, originally proposed by Diggle and Kenward (1994), which assumes that drop-out depends on the entire future only through the next observation. But in a previous paper (Rotnitzky et al., 1998, p. 1321), we showed that this assumption is itself implausible.

To clarify matters, we adopt LR's discrete time model in which Q is the censoring time and the complete data are (Y_1, Y_2, Y_3). For a subject who is censored in the interval $(1, 2], Q = 1$; for one censored in $(2, 3], Q = 2$; and for an uncensored subject, $Q = 3$. LR claim that the discrete time analogue of our model implies $1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = (1 - \text{pr}(Q = 1|Y_1))\Psi(\alpha_0 Y_3)$ and that a more "plausible" model would be $1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = (1 - \text{pr}(Q = 1|Y_1))\Psi(\alpha_0 Y_2)$, where $\Psi(u) = \exp\{-\exp(u)\}$ corresponds to the complementary log-log link function. However, LR's claims are incorrect. Indeed, there does not exist any joint distribution for (Q, Y_1, Y_2, Y_3) that satisfies either of these equations. To see this note that by taking conditional expectations with respect to Y_1 on both sides of the first equation, one deduces that $E\{\Psi(\alpha_0 Y_3)|Y_1\} = 1$ with probability one. But this is impossible because the range of the function $\Psi(u)$ is the open interval (0, 1). An identical argument applied to the second equation leads to $E\{\Psi(\alpha_0 Y_2)|Y_1\} = 1$ which again cannot occur.

LR's misstatements can be easily corrected and, once corrected, do not materially affect the thrust of their argument which we still find to be misguided. Specifically, the correct discrete time analogue of our model (1) has $1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = \Psi\{h(Y_1) + \alpha_0 Y_3\}$ where $h(Y_1)$ is an unspecified nuisance function, and α_0 is regarded as known. The correct form of LR's more "plausible" model has $1 - \text{pr}(Q = 1|Y_1, Y_2, Y_3) = \Psi\{h(Y_1) + \alpha_0 Y_2\}$ where $h(Y_1)$ is an unspecified function of Y_1. LR's parametric model imposes the additional assumption that $h(Y_1)$ is linear in Y_1. The Diggle–Kenward model only differs from LR's parametric model in that the complementary log-log link is replaced by the logistic link. Finally, the discrete time version of model (15) is $1 - \text{pr}(Q = 1|Y_1, Y_3) = \Psi\{h(Y_1) + \alpha_0 Y_3\}$ with $h(Y_1)$ unspecified and α_0 regarded as known. Subsequent discussion refers to the correct discrete time versions of these models. Both model (1) and the LR–Diggle–Kenward model impose implausible conditional in-

dependence assumptions: Conditional on the past Y_1, drop-out in the interval (1, 2] depends on the future (Y_2, Y_3) only through the final observation in Y_3 in (1) and only through the next observation Y_2 in LR–Diggle–Kenward. In contrast, (15) imposes no conditional independence assumptions; the fact that the right side of the equation (15) does not depend on Y_2 is not an assumption, but rather a simple logical consequence of the fact that (15) models the discrete hazard $\text{pr}(Q = 1|Y_1, Y_3)$ of Q at $t = 1$ given the past Y_1 and the final observation Y_3.

When we are interested in performing a sensitivity analysis over the magnitude of the final outcome's influence on selection, (15) is a good choice, because it lessens the difficulty of choosing a plausible selection function. Indeed, we introduced (1) and (13) before (15) only for pedagogic reasons. Specifically, a theoretical analysis of the semiparametric model (15) is both less familiar and more subtle than the analysis of (13). Remarkably, however, we show in Section 7.3.1 and in Appendix A that the interval estimators for (13) remain valid when (15) is true but (13) is false. A better understanding of (15) can be obtained by reviewing the special case in which $\alpha_0 = 0$ studied by Robins et al. (1995), which we do in Section 3.3 here.

Finally, to review why the conditional independence assumption imposed by the LR–Diggle–Kenward model is implausible, we note that it will be false under the following, quite natural scenario. Given the data up to time j, (a) the hazard of drop-out given in the interval $(j, j+1]$ is a deterministic function of some unmeasured vector-valued latent variable $\mathbf{U}_{j+1} = (U_{j+1,1}, \ldots, U_{j+1,k})$ encoding multiple aspects of a subject's health, economic, and emotional status at that time; (b) the $\mathbf{U}_j$ are highly correlated over time; and (c) Y_{j+1} is a possibly mismeasured version of some component of $\mathbf{U}_{j+1}$. Indeed, Little (1995, p. 1116) himself previously noted that the Diggle–Kenward model could not be true if the Y_j's were a mismeasured version of an underlying $\mathbf{U}_j$ that determined drop-out.

2.5.5 Censoring by Death. LR follow Robins (1995, p. 249) in arguing that censoring due to death should often be handled differently from censoring due to other causes. We agree that it would often be wise to impose the assumption that a subject's CD4 count at end of follow-up T is not defined if he or she died prior to T. Robins (1995, p. 249) discussed the strong nonidentifiable restrictions necessary to identify even the null hypothesis of no treatment effect under this assumption. In our article we did not make this assumption, because it would have complicated the analysis and drawn attention away from our main points.

2.6 Laird and Pauler

2.6.1 Nonparametric Versus Saturated Models. Baker and Laird (1988) demonstrated that with categorical data, a nonignorable model may fail to exactly fit the data (i.e., the expected cell counts may differ from the observed cell counts), even though the model is "saturated" in the sense that the number of free parameters in the model equals the total degrees of freedom. LP incorrectly suggest that the discrete time, categorical data versions of our models studied by Rotnitzky et al. (1998, sec. 7) might suffer from similar lack of fit. LP failed to recognize that a model that is saturated in the number of parameters is not necessarily nonparametric. Recall that a model for missing data is nonparametric if, for each possible law F_O for the observed data, there is a joint law allowed by the model whose marginal is exactly F_O. Our models (1), (13), (15), and (16) are nonparametric models. With categorical data, the MLE of the expected cell counts under a nonparametric model will always equal the observed cell counts.

It is often argued that with categorical missing data, a model that imposes only the MAR assumption is nonparametric, because the number of free parameters in an unrestricted MAR model equals the degrees of freedom. As the above discussion makes clear, this argument is flawed. Gill, van der Laan, and Robins (1997), however, provided a rigorous proof that a model for categorical missing data that imposes only MAR is nonparametric. The Gill et al. proof covers both monotone and nonmonotone missing data.

2.6.2 Laird and Pauler's Example 2. In their Example 2, LP suggest considering the discrete time version of our mean model (16), which they refer to as a pattern-mixture model. However, strictly speaking, model (16) is not a pattern mixture model in the sense of Little (1993a). In our paper, we therefore referred to model (16) as a "sequential pattern-mixture model." LP assume that censoring by Q occurs only at times $t \in \{8, 20, 32, 44\}$, $\mathbf{V}(t)$ is the last observed value of $\mathbf{V}$ when $Q = t, Y$ is measured at time $T = 56$, and $Q \equiv T$ if Y is uncensored. Further they suggest assuming that the difference between the conditional mean $E[Y|\bar{\mathbf{V}}(t), Q = t]$ of Y among those dropping out at t and the mean $E[Y|\bar{\mathbf{V}}(t), Q > t]$ of those continuing on study is a constant ϕ, which is regarded as known but then varied in a sensitivity analysis. They appear to suggest fitting this model by further specifying a regression model for $E[Y|\bar{\mathbf{V}}(t), Q > t]$. In section 3.5.2 here we show that LP's suggested approach can be used to construct a consistent estimator of μ provided the regression model for $E[Y|\bar{\mathbf{V}}(t), Q > t]$ is correct.

When $\phi = 0$, we show in Section 3.5.2 below that LP's suggested approach reduces to the iterated conditional expectations estimator (ICE) studied by Robins et al. (1995) and Robins (1998). Suppose that the specified model for $E[Y|\bar{\mathbf{V}}(t), Q > t]$ is non-linear in $\bar{\mathbf{V}}(t)$. Consider the special case in which the data happens to be MCAR (which implies $\phi = 0$). Even in this special case the ICE estimator will be inconsistent unless the regression model for $E[Y|\bar{\mathbf{V}}(t), Q > t]$ happens to be correct; in contrast, the estimators we propose in section 7.3.2 of our paper and Section 3.5.2 below are guaranteed to be consistent.

A more fundamental problem with LP's suggested approach is that, as discussed in Robins et al. (1995) and section 3.5.2 below, the four nonlinear regression models $m_t(\bar{\mathbf{V}}(t), \beta_t)$ for $E[Y|\bar{\mathbf{V}}(t), Q > t], t = 8, 20, 32, 44$, will almost always be incompatible in the sense that for nearly all values of the parameters β_t, there will exist no joint distri-

bution of $(Q, \bar{\mathbf{V}}(T), Y)$ that simultaneously satisfies all four models.

As an alternative to our model (16), LP also suggest relating the mean of Y of those who drop-out at time t with that of those that drop-out at time $t+1$, by, for example, specifying $E[Y|\bar{\mathbf{V}}(t), Q=t] = E[Y|\bar{\mathbf{V}}(t), Q=t+1] + \phi_1$, or by relating the mean of Y of those who drop-out at t to that of the completers $(Q=T)$, by, for example, specifying $E[Y|\bar{\mathbf{V}}(t), Q=t] = E[Y|\bar{\mathbf{V}}(t), Q=T] + \phi_2$. Conducting a sensitivity analysis based on these models may be unsatisfactory because: (i) as emphasized by Robins (1998) and Little and Rubin in their discussion here, one would nearly always wish the model to include the possibility that one has succeeded in collecting in $\mathbf{V}(t)$ data on all the important causes of drop-out so that the data are missing at random (MAR), i.e. $\text{pr}(Q=t|Q \geq t, \bar{\mathbf{V}}(T), Y) = \text{pr}(Q=t|Q \geq t, \bar{\mathbf{V}}(t))$ and yet, (ii) MAR does not imply that ϕ_1 or ϕ_2 take any particular fixed value such as 0. In contrast, MAR implies $\phi = 0$ in model (16).

3. MORE TECHNICAL DISCUSSION

In the remainder of this rejoinder we treat various issues in greater depth. An understanding of this discussion requires familiarity with the terminology and notation that we and coauthors have used in a series of articles whose purpose has been to generalize the results of Heitjan and Rubin (1991) and Rubin (1976) for parametric missing-data and censored-data models to include parametric, semiparametric, and nonparametric models.

3.1 Review of Notation and Terminology

Let L denote a subject's full (possibly incompletely observed) data vector. Assume that there are available for data analysis n iid copies of the observed data $O = (R, c_R(L))$, where R is a random vector and $c_R(L)$ is a known function; that is, a coarsening of L that changes with R. The coarsening variable R indicates what part of L is observed. Rubin (1976) referred to $c_R(L)$ as L_{obs}. Let $\Delta = I[c_R(L) = L]$ be the indicator of observing full (i.e., complete) data. Coarsened data includes missing data and censored data, as well as other partially observed data configurations. For example, missing data is the special case of coarsened data in which each (possibly multivariate) component L_k of $L = (L_1', \ldots, L_p')'$ is observed either exactly or not at all and $R = (R_1, \ldots, R_p)'$ is a vector of response indicators; that is, L_k is observed if and only if $R_k = 1$. Right-censored failure time data with a time-varying covariate process $\mathbf{V}(t)$ is coarsened data in which $O = (X = \min(\mathcal{T}, Q), \Delta = I(X = \mathcal{T}), \bar{\mathbf{V}}(X)), L = (\bar{\mathbf{V}}(\mathcal{T}), \mathcal{T}), \mathcal{T}$ is a failure time random variable, $\bar{\mathbf{V}}(t) = \{\mathbf{V}(u); 0 \leq u \leq t\}, R$ is equal to the censoring time Q if $Q < \mathcal{T}$, and, by convention, $R = \infty$ if $\mathcal{T} \leq Q$. Also, $c_r(L)$ is the event $(\mathcal{T} > r, \bar{\mathbf{V}}(r))$ when $r < \infty$ and $c_\infty(L) = L$ (Robins and Rotnitzky 1992). Our article's data structure is the special case in which $\mathcal{T}$ is equal to the nonrandom end of follow-up time T with probability 1 and $\mathbf{V}(T) = Y$.

A model for the joint distribution of (R, L) specifies that $f(r,l) \in \{f(r,l;\rho), \rho \in \boldsymbol{\rho}\}$, where $f(r,l)$ is the true joint density, $\boldsymbol{\rho}$ is a possibly infinite-dimensional parameter space, and $f(r,l;\rho)$ is a known density with respect to a dominating measure. The model is correctly specified if the true density $f(r,l)$ equals $f(r,l;\rho_0)$ for some $\rho_0 \in \boldsymbol{\rho}$. A model $f(r,l;\rho)$ is a selection model if $f(r,l;\rho) = f(r|l;\rho_{\text{mis}})f(l;\rho_{\text{ful}})$, and the parameters ρ_{mis} and ρ_{ful} are variation-independent; that is, $\rho_{\text{mis}} \in \boldsymbol{\rho}_{\text{mis}}, \rho_{\text{ful}} \in \boldsymbol{\rho}_{\text{ful}}$, and $\boldsymbol{\rho} = \boldsymbol{\rho}_{\text{mis}} \times \boldsymbol{\rho}_{\text{ful}}$. The model is parametric if $\boldsymbol{\rho}$ can be smoothly identified with a subset of a finite-dimensional Euclidean space, in which case we refer to ρ as a finite-dimensional parameter; otherwise, the model is said to be semiparametric or, equivalently, infinite dimensional. An estimator $\hat{\mu}$ of a finite dimensional functional $\mu \equiv \mu(\rho)$ of ρ is a consistent and asymptotically normal (CAN) estimator at a given $\rho \in \boldsymbol{\rho}$ if $n^{1/2}(\hat{\mu} - \mu(\rho))$ converges in law to a $N(0, \sigma^2)$ distribution. The estimator is regular CAN (RCAN) at ρ if $n^{1/2}(\hat{\mu} - \mu(\rho_n))$ converges to the same $N(0, \sigma^2)$ distribution under all sequences $\{\rho_n\}$ such that $n^{1/2}(\rho_n - \rho)$ is bounded and $\{\rho_n\}$ is contained in a regular parametric submodel of $\boldsymbol{\rho}$. In a parametric model containing ρ, a necessary condition for a RCAN estimator $\hat{\mu}$ of μ to exist at ρ is that the Cramer–Rao variance bound for μ at ρ is finite. In a semiparametric model a necessary condition is that the semiparametric variance bound for μ at ρ is finite (Bickel et al. 1993). The semiparametric variance bound for μ at a law $\rho \in \boldsymbol{\rho}$ is the supremum of the parametric Cramer–Rao variance bounds for μ at ρ over all parametric submodels containing ρ. RCAN estimators may exist at some but not all laws allowed by a model. Although regularity is not often explicitly mentioned, it is usually implicitly assumed. For example, in a parametric model the Cramer–Rao bound is the minimal asymptotic variance that can be obtained by a RCAN estimator; nonregular CAN estimators may have smaller asymptotic variance.

We focus our discussion on semiparametric selection models in which $\rho_{\text{ful}} = (\mu, \theta)$, where μ is a finite- (say p) dimensional parameter of interest and θ is an infinite-dimensional nuisance parameter. For simplicity, until Section 3.2.8, we suppose that μ is a one-dimensional functional such as the mean or median of a component Y of L and the model $f(l;\rho_{\text{ful}})$ places no restriction on the law of L. This implies that in the absence of missing data, all RCAN estimators of μ are, up to asymptotic equivalence, equal to solutions to a particular unbiased estimating equation $\sum_i m(L_i;\mu) = 0$ (Bickel et al. 1993). [The last statement is technically true only for regular asymptotically linear (RAL) estimators, a subset of RCAN estimators. But because RCAN estimators that are not RAL are rather pathological and unlikely to arise in applications, we ignore the distinction between RCAN and RAL estimators.] As an example, $m(L;\mu)$ is $Y - \mu$ for the mean.

When data on Y are missing for some subjects and $f(r|l)$ is unrestricted, μ is not identified. Thus it is natural to consider more restrictive models in which $\rho_{\text{mis}}, \rho_{\text{ful}}$ (equivalently θ), or both lie in finite-dimensional parameter sets that we call parametric nonresponse models, parametric complete-data models, and (fully) parametric models. For notational convenience, we let R1 denote a parametric nonresponse model, R2 denote a parametric complete-

data model, and R3 denote a (fully) parametric model. One might expect a large degree of symmetry between models R1 and R2. Surprisingly, this is not the case. Specifically, no RCAN estimator of μ (and of ρ_{ful}) exist in any parametric complete-data model R2 in which $f(r|l)$ is unrestricted (Rotnitzky, et al. 1998). In contrast, RCAN estimators of μ often exist at most $\rho \in \boldsymbol{\rho}$ in parametric nonresponse models R1 even though $f(l)$ is unrestricted. Rotnitzky and Robins [1997, eqs. (29) and (30)] provided necessary conditions for their existence, and Rotnitzky et al. (1998) studied many examples. A semiparametric nonresponse model (model R0) differs from model R1 in that $\rho_{\text{mis}} = (\gamma, \eta)$ where γ is finite (say q) dimensional and η is infinite dimensional. Model $B(\alpha_0)$ of our article is an example with η being the cumulative baseline hazard function $\Lambda_0(\cdot)$.

We say that there is positive probability of observing complete data if the chance $\text{pr}(\Delta = 1|L)$ of fully observing L is always positive; that is,

$$\pi(L) \equiv \text{pr}(\Delta = 1|L) > 0 \quad \text{with probability 1.} \tag{1}$$

Unless we state otherwise, assume that (1) holds. Then any RCAN estimator of $\psi = (\mu, \rho'_{\text{mis}})', \mu \in R^1, \rho_{\text{mis}} \in R^q$ in the parametric nonresponse model R1 is asymptotically equivalent to an AIPCW estimator $\hat{\psi} \equiv \hat{\psi}(\hat{b}) = (\hat{\mu}(\hat{b}), \hat{\rho}_{\text{mis}}(\hat{b})')'$ solving

$$\sum_i h(O_i; \mu, \rho_{\text{mis}}; \hat{b}) = 0,$$

$$\begin{aligned} &h(O; \mu, \rho_{\text{mis}}; b) \\ &\quad = \Delta\pi^{-1}(L; \rho_{\text{mis}})\{d(L; \mu) - E_{\rho_{\text{mis}}}[(1-\Delta)b(O; \psi)|L]\} \\ &\qquad + (1-\Delta)b(O; \psi) \end{aligned} \tag{2}$$

for some, possibly data-dependent $(1 + q)$-dimensional function $\hat{b}$ chosen by the analyst that converges in probability as n goes to ∞ to a fixed function. Here $d(L; \mu) = (m(L; \mu), \mathbf{0}')'$ and $\pi^{-1}(L; \rho_{\text{mis}}) = 1/\pi(L; \rho_{\text{mis}})$ (Rotnitzky and Robins 1997). Any RCAN estimator of $\psi = (\mu, \gamma) \in R^{1+q}$ in the semiparametric nonresponse model R0 is asymptotically equivalent to an estimator $\hat{\psi}(\hat{b}) = (\hat{\mu}(\hat{b}), \hat{\gamma}(\hat{b})')'$ provided that in (2) we replace ρ_{mis} by $\hat{\rho}_{\text{mis}}(\gamma) \equiv (\gamma, \hat{\eta}(\gamma))$ and $\hat{\eta}(\gamma)$ converges to the true value of η at an appropriate rate. As an example, the estimator $\hat{\psi}(\hat{b})$ of Section 4 of our article is a RCAN estimator in model $B(\alpha_0)$, where $\eta = \Lambda_0(\cdot), \hat{\eta}(\gamma) = \hat{\Lambda}(\cdot; \gamma)$.

Efficiency Theory. Let $\hat{\rho}_{\text{MLE}} = (\hat{\rho}_{\text{ful,MLE}}, \hat{\rho}_{\text{mis,MLE}})$ be the MLE of ρ under a parametric model R3. In model R1, Rotnitzky and Robins (1997) showed that the unique efficient non-random choice $b_{\text{eff}}(\cdot; \psi)$ of $\hat{b}(\cdot; \psi)$ is a function of all of the components of the ρ generating the data. One possibility is to evaluate this function at $\hat{\rho} = (\psi, \hat{\theta})$ where $\hat{\theta}$ is an estimate of θ based on a parametric submodel $f(l; \rho_{\text{ful}})$. The resulting estimator $\hat{\psi}_{\text{loc,eff}} = \hat{\psi}(\hat{b}_{\text{eff}})$ is locally semiparametric efficient at the chosen parametric submodel $f(l; \rho_{\text{ful}})$. That is, (a) $\hat{\psi}_{\text{loc,eff}}$ is a RCAN estimator under model R1 even if the parametric submodel $f(l; \rho_{\text{ful}})$ is misspecified and (b) $\hat{\psi}_{\text{loc,eff}}$ is the most efficient estimator satisfying (a) when both model R1 and the submodel $f(l; \rho_{\text{ful}})$ are correct. When the submodel $f(l; \rho_{\text{ful}})$ is correct, $\hat{\mu}_{\text{loc,eff}}$ has asymptotic variance equal to the semiparametric variance bound. Thus in model R1, $\hat{\mu}_{\text{loc,eff}}$ has an asymptotic variance greater than or equal to that of the MLE $\hat{\mu}_{\text{MLE}} = \int y\, dF(l; \hat{\rho}_{\text{ful,MLE}})$ under the associated fully parametric submodel of model R1. If, however, $f(l; \rho_{\text{ful}})$ is misspecified but model R1 is correct, $\hat{\mu}_{\text{MLE}}$, in contrast to $\hat{\mu}_{\text{loc,eff}}$, is inconsistent. To make this discussion less abstract, we use, as requested by LR, a simple missing-data structure as a running example.

Example. Suppose that $L = (Y, \mathbf{V}), \mathbf{V}$ is an always observed high-dimensional vector of baseline variables (with all components continuous) and Y, whose mean μ is the parameter of interest, may be missing for some subjects. Thus $O = (\Delta, L_{\text{obs}})$, where $\Delta = R, L_{\text{obs}} = (Y, \mathbf{V})$ when $\Delta = 1$ and $L_{\text{obs}} = \mathbf{V}$ when $\Delta = 0$. We consider the parametric nonresponse model R1 with $f(r|l; \rho_{\text{mis}})$ given by

$$\begin{aligned} &\text{logit pr}[\Delta = 1|L; \rho_{\text{mis}}] \\ &\qquad \equiv \text{logit } \pi(L; \rho_{\text{mis}}) = \gamma_0 + \gamma_1'\mathbf{V} + \alpha Y, \end{aligned} \tag{3}$$

where $\rho_{\text{mis}} = (\gamma', \alpha)'$. Neither μ nor α need be identified. For example, suppose the true but unknown value of α is 0. Then it is not difficult to show that μ is *not* identified from the law of F_O of O if and only if the conditional moment generating function MGF $\log E[\exp\{tY\}|\Delta = 1, \mathbf{V}]$ exists and is a linear function of $\mathbf{V}$ for some non zero-value of t. Further it follows from Rotnitzky and Robins (1997, Table 2, Model 5), that a RCAN estimator cannot exist whenever $E[Y|\mathbf{V}, \Delta = 1]$ is linear in $\mathbf{V}$. Hence when the conditional MGF does not exist for all $t \neq 0$, the unknown α is zero, and $E[Y|\mathbf{V}, \Delta = 1]$ is linear in $\mathbf{V}, \mu$ is identified but no RCAN estimator exists and therefore the rate of convergence of the optimal estimator will be less than the usual $n^{1/2}$. If, however, as when conducting a sensitivity analysis, we consider the submodel in which we regard α as known, so that $\rho_{\text{mis}} = \gamma$, then RCAN estimators exist at all laws allowed by the submodel.

In this simple example, to obtain $\hat{\mu}_{\text{loc,eff}}$ in model (3) at a parametric submodel $f(l; \rho_{\text{ful}})$, it is sufficient to evaluate $\hat{b}_{\text{eff}}(O; \psi)$ for subjects with incomplete data ($\Delta = 0$), which Rotnitzky and Robins (1997) showed to be $\hat{b}_{\text{eff}}(O; \psi) = E^*_{\hat{\rho}_{r\text{MLE}}(\psi)}[H(\psi)|\mathbf{V}]$, where $H(\psi) = \pi(L; \rho_{\text{mis}})[\pi^{-1}(L; \rho_{\text{mis}})(Y - \mu), 1, \mathbf{V}', Y]'$, $E^*_\rho[H|\mathbf{V}] \equiv E_{\rho_{\text{ful}}}[H\{\pi^{-1}(L; \rho_{\text{mis}}) - 1\}|\mathbf{V}]/E_{\rho_{\text{ful}}}[\pi^{-1}(L; \rho_{\text{mis}}) - 1|\mathbf{V}]$, and $\hat{\rho}_{r\text{MLE}}(\psi)$ is the restricted MLE of ρ in the parametric submodel R3 with $\psi = (\mu, \rho_{\text{mis}})$ held fixed.

3.2 Parametric Versus Semiparametric Inference

In this section we compare parametric and semiparametric inference. We require the following definitions. If the conditional density $f(R|L)$ of R given L is only a function of O, we say that $f(R|L)$ satisfies coarsening at random (CAR) (Heitjan and Rubin 1991). In our example the law $f(R|L)$ of (3) satisfies CAR if and only if $\alpha = 0$. Any missing-data model is said to be a CAR model if and only if

$f(r, l; \rho)$ satisfies CAR for all $\rho \in \boldsymbol{\rho}$. A CAR selection model is ignorable in the sense of Rubin (1976) for the purposes of frequentist inference. It is ignorable for Bayesian inference if ρ_{mis} and ρ_{ful} are a priori independent. If a selection model includes any density $f(r|l; \rho_{\text{mis}})$ that fails to satisfy CAR, then the model is nonignorable. In a CAR parametric nonresponse model, the MLE $\hat{\rho}_{\text{mis}}$ of ρ_{mis} equals $\hat{\rho}_{\text{mis,loc,eff}}$.

Example (Continued). Model (3) is a nonignorable selection model because it includes nonignorable laws ($\alpha \neq 0$) in addition to ignorable laws ($\alpha = 0$). As an example of a CAR model, consider the submodel of (3) in which, as when conducting a sensitivity analysis, we regard α as known, so $\rho_{\text{mis}} = \gamma$. Then this submodel is a CAR model if and only if $\alpha = 0$. In that case, $\pi(L; \rho_{\text{mis}})$ depends only on $\mathbf{V}$, and so $E^*_{\rho}[H|\mathbf{V}]$ simplifies to $E_{\rho_{\text{ful}}}[H|\mathbf{V}]$.

We show that the parametric MLE, $\hat{\mu}_{\text{MLE}}$ is less robust to model misspecification than (a) any AIPCW estimator $\hat{\mu}(\hat{b})$ in a nonignorable selection model and (b) a locally efficient AIPCW estimator in a ignorable selection model. Before proceeding, we offer a note of caution. Many of the arguments made in this section and in our article rely on large-sample theory and thus may not be relevant to studies with small sample sizes. When there is doubt, investigation by simulation would be warranted.

3.2.1 Robustness of the Augmented Inverse Probability of Censoring Weighted Estimators $\hat{\mu}(b)$ in Nonignorable Selection Models. In a nonignorable parametric selection model R3, the MLE $\hat{\mu}_{\text{MLE}}$ will generally be inconsistent unless the parametric models $f(r|l; \rho_{\text{mis}})$ and $f(l; \rho_{\text{ful}})$ are both correctly specified. Somewhat surprisingly, this remains the case even if the true law $f(r|l)$ satisfies CAR. In contrast, $\hat{\mu}(b)$ generally will be CAN if the model $f(r|l; \rho_{\text{mis}})$ is correct.

3.2.2 Robustness of $\hat{\mu}(b)$ in Designed Studies. In studies, such as sample surveys, with missingness only by design, $f(r|l)$ is known, so that ρ_{mis} need not be estimated and its dimension q can be taken to be 0 in model R1. It follows that any estimator $\hat{\mu}(\hat{b})$ is guaranteed to be a RCAN estimator of μ in model R1. In contrast, $\hat{\mu}_{\text{MLE}}$ will be inconsistent if the parametric model $f(l; \rho_{\text{ful}})$ is misspecified.

3.2.3 Robustness of $\hat{\mu}_{\text{loc,eff}}$ in Coarsened at Random Selection Models. In CAR selection models, $\hat{\mu}_{\text{MLE}}$ is CAN if the parametric model is correct and the true density $f(r|l)$ satisfies CAR, even if the parametric model $f(r|l; \rho_{\text{mis}})$ is misspecified. On the other hand, the estimators $\hat{\mu}(b)$ are CAN whenever the parametric CAR model $f(r|l; \rho_{\text{mis}})$ is correct, even when $f(l; \rho_{\text{ful}})$ is misspecified. Suppose that, as is the case in studies with unplanned missing data, we cannot be certain that either parametric model is correct. We still may hope to find estimators of μ that are CAN if either of the two models is correct and $f(r|l)$ satisfies CAR. The estimator $\hat{\mu}_{\text{MLE}}$ fails to satisfy this goal, because it is inconsistent if the parametric model $f(l; \rho_{\text{ful}})$ is misspecified. However, following Robins and Ritov (1997) and Robins, Rotnitzky, and van der Laan (to appear), we show in Section 3.2.9 that the estimator $\hat{\mu}_{\text{loc,eff}}$ satisfies this goal. Thus, somewhat surprisingly, even in ignorable models locally efficient AIPCW estimators are more robust than parametric MLEs.

Example (Continued): Further Robustness and Regression-Propensity Score Estimators in Ignorable Models. We can sometimes further increase the robustness of $\hat{\mu}_{\text{loc,eff}} = \hat{\mu}(\hat{b}_{\text{eff}})$ by choosing $\hat{b}_{\text{eff}}$ differently. A parametric model $e(\mathbf{V}; \omega)$ for the regression function $E[Y|\mathbf{V}]$ is less restrictive than a parametric model $f(l; \rho_{\text{ful}})$ for the joint law of $L = (Y, \mathbf{V})$, because a correct model for $f(l)$ implies a correct model for $E[Y|\mathbf{V}]$, but not conversely. Hence if we choose $\hat{b}_{\text{eff}}(\mathbf{V}; \mu) = e(\mathbf{V}; \hat{\omega}) - \mu$ with $\hat{\omega}$ the (possibly nonlinear) least squares estimator of the parameter ω among subjects with $\Delta = 1$, then $\hat{\mu}(\hat{b}_{\text{eff}}) = n^{-1}\{\sum_i e(\mathbf{V}_i; \hat{\omega}) + \Delta_i \pi^{-1}(\mathbf{V}_i; \hat{\rho}_{\text{mis}})[Y_i - e(\mathbf{V}_i; \hat{\omega})]\}$ will be CAN if either the model $e(\mathbf{V}; \omega)$ or the model $f(r|l; \rho_{\text{mis}})$ is correct. Further, if both are correct, then the asymptotic variance will equal the semiparametric variance bound for model R1. Consider now the special case in which $e(\mathbf{V}; \omega) \equiv e(\mathbf{V}; \omega, \hat{\rho}_{\text{mis}})$ is $\Phi\{s(\mathbf{V}; \omega_1) + \omega_2 \pi^{-1}(\mathbf{V}; \hat{\rho}_{\text{mis}})\}$, where $s(\mathbf{V}; \omega_1)$ is a known function, $w = (w_1', w_2)', \Phi^{-1}$ is a known link function, and $\hat{\omega}$ solves $0 = \sum_i \Delta_i \{\partial s(\mathbf{V}_i; \omega_1)/\partial \omega_1', \pi^{-1}(\mathbf{V}_i; \hat{\rho}_{\text{mis}})\}'\{Y_i - e(\mathbf{V}_i; \omega)\}$. Then $\hat{\mu}(\hat{b}_{\text{eff}}) = n^{-1} \sum_i e(\mathbf{V}_i; \hat{\omega})$, because the terms in Δ_i exactly cancel. The term $\omega_2 \pi^{-1}(\mathbf{V}; \hat{\rho}_{\text{mis}})$ protects against misspecification of the model $e(\mathbf{V}; \omega, \rho^*_{\text{mis}})$ for $E[Y|\mathbf{V}]$, where ρ^*_{mis} is the probability limit of the MLE $\hat{\rho}_{\text{mis}}$. If Φ^{-1} is a canonical link function of a generalized linear model, then $\hat{\omega}$ is the iterately reweighted least squares (IRLS) estimator solving the quasi-likelihood score equation. The representation of a locally efficient AIPCW estimator as the regression estimator $n^{-1} \sum_i e(\mathbf{V}_i; \hat{\omega})$ should help dispel the mistaken belief that all AIPCW estimators are inefficient.

Extension to Estimation of Treatment Effects. A straightforward generalization of this estimator solves the longstanding problem in the analysis of treatment effects of how to add the propensity score to a regression model to guarantee consistency, without needing to smooth. Consider an observational study with n iid copies of $(\Delta, Y, \mathbf{V})$ with Δ the dichotomous treatment indicator, Y the outcome, $\mathbf{V}$ a vector of pretreatment variables, $\pi(\mathbf{V}; \rho_{\text{mis}})$ a parametric model for the propensity score $\text{pr}(\Delta = 1|\mathbf{V})$ with MLE $\hat{\rho}_{\text{mis}}$ and probability limit ρ^*_{mis}, and $\Phi\{s(\Delta, \mathbf{V}; \omega_1)\}$ a canonical link generalized linear model for $E[Y|\Delta, \mathbf{V}]$. Assume conditional ignorability of treatment given $\mathbf{V}$; that is, $Y(\delta) \coprod \Delta|\mathbf{V}$, where $Y(\delta)$ is the possibly counterfactual outcome at treatment level δ. Then the average treatment effect is $\mu \equiv E\{E[Y|\Delta = 1, \mathbf{V}] - E[Y|\Delta = 0, \mathbf{V}]\}$. Let $\hat{\omega}$ be the IRLS estimate in the canonical link expanded model $e(\Delta, \mathbf{V}; \omega) \equiv e(\Delta, \mathbf{V}; \omega, \hat{\rho}_{\text{mis}}) \equiv \Phi[s(\Delta, \mathbf{V}; \omega_1) + \omega_2 \Delta \pi^{-1}(\mathbf{V}; \hat{\rho}_{\text{mis}}) + \omega_3(1 - \Delta)\{1 - \pi(\mathbf{V}; \hat{\rho}_{\text{mis}})\}^{-1}]$ for $E[Y|\Delta, \mathbf{V}]$. Then $\hat{\mu} = n^{-1} \sum_i e(1, \mathbf{V}_i; \hat{\omega}) - e(0, \mathbf{V}_i; \hat{\omega})$ is a CAN estimator of μ if either the model $e(\Delta, \mathbf{V}; \omega, \rho^*_{\text{mis}})$ or $\pi(\mathbf{V}; \rho_{\text{mis}})$ is correct and is locally semiparametric efficient at submodel $e(\Delta, \mathbf{V}; \omega, \rho^*_{\text{mis}})$ in the semiparametric model characterized by model $\pi(\mathbf{V}; \rho_{\text{mis}})$. To guaran-

tee consistency of $\hat{\mu}$ under misspecification of $s(\Delta, \mathbf{V}; \omega_1)$, it was necessary to add both the terms $\Delta\pi^{-1}(\mathbf{V}; \hat{\rho}_{\text{mis}})$ and $(1-\Delta)\{1-\pi(\mathbf{V}; \hat{\rho}_{\text{mis}})\}^{-1}$.

Suppose now that Δ is a possibly multivariate continuous and/or discrete treatment. We specify a parametric model $f(\Delta|\mathbf{V}; \rho_{\text{mis}})$ for the conditional density of Δ given $\mathbf{V}$ with respect to a dominating measure $\nu(\Delta)$. Let $\hat{\rho}_{\text{mis}}$ be the MLE with probability limit ρ^*_{mis}. We also specify a marginal structural mean (MSM) model $g(\delta, \mathbf{w}; \mu)$ for $E[Y(\delta)|\mathbf{W} = \mathbf{w}]$, where $\mathbf{W}$ is a subvector of $\mathbf{V}$, $g(\delta, \mathbf{w}; \mu)$ is a known function, and μ is an unknown parameter vector (Robins, 1999; Robins, Greenland, and Hu, 1999). Finally, we specify a canonical link working model $\Phi[s(\Delta, \mathbf{V}; \omega_1)]$ for $E[Y|\Delta, \mathbf{V}]$. Then, given a user-supplied function $m(\Delta, \mathbf{W})$ of the dimension of μ, we let $\tilde{\mu}$ solve $\sum_i \int [m(\Delta, \mathbf{W}_i)\{e(\Delta, \mathbf{V}_i; \hat{\omega}) - g(\Delta, \mathbf{W}_i; \mu)\}] d\nu(\Delta) = 0$ where $e(\Delta_i, \mathbf{V}_i; \hat{\omega})$ and $\hat{\omega}$ are the predicted value of Y_i and the IRLS estimator in the expanded working model $e(\Delta, \mathbf{V}; \omega) \equiv e(\Delta, \mathbf{V}; \omega, \hat{\rho}_{\text{mis}}) \equiv \Phi[s(\Delta, \mathbf{V}; \omega_1) + \omega_2' m(\Delta, \mathbf{W})/f(\Delta|\mathbf{V}; \hat{\rho}_{\text{mis}})]$. Then given that both model $g(\delta, \mathbf{w}; \mu)$ and conditional ignorability given $\mathbf{V}$ hold, $\tilde{\mu}$ is a CAN estimator if at least one of the models $e(\Delta, \mathbf{V}; \omega, \rho^*_{\text{mis}})$ or $f(\Delta|\mathbf{V}; \rho_{\text{mis}})$ is correct. Furthermore, in the semiparametric model characterized by the MSM model $g(\delta, \mathbf{w}; \mu)$, conditional ignorability, and the model $f(\Delta|\mathbf{V}; \rho_{\text{mis}})$, for a certain choice $m_{\text{eff}}(\Delta, \mathbf{W})$ of $m(\Delta, \mathbf{W})$, $\hat{\mu}$ is locally efficient at the submodel $e(\Delta, \mathbf{V}; \omega, \rho^*_{\text{mis}})$. Robins (1999) shows how to obtain $m_{\text{eff}}(\Delta, \mathbf{W})$.

Model $e(\Delta, \mathbf{V}; \omega, \rho^*_{\text{mis}})$ for $E[Y|\Delta, \mathbf{V}]$ will often be incompatible with the MSM model $g(\delta, \mathbf{w}; \mu)$. One possible solution is to define the model $g(\delta, \mathbf{w}; \mu)$ in terms of the model $e(\Delta, \mathbf{V}; \omega, \rho^*_{\text{mis}})$ via $g(\delta, \mathbf{W}; \mu) \equiv g(\delta, \mathbf{W}; \mu, \hat{\rho}_{\text{mis}}) \equiv \int e(\delta, \mathbf{V}; \omega, \hat{\rho}_{\text{mis}}) dF(\mathbf{V}|\mathbf{W}; \kappa)$, where $\mu = (\omega, \kappa)$ and $f(\mathbf{V}|\mathbf{W}; \kappa)$ is a model for $f(\mathbf{V}|\mathbf{W})$. Estimation of μ by $\tilde{\mu}$ is as described above. Both the estimators $\hat{\mu}$ and $\tilde{\mu}$ are special cases of the general class of augmented inverse probability of treatment weighted (IPTW) estimators of marginal structural mean (MSM) models proposed by Robins (1999). Robins (2000) extends these results by deriving a "regression estimator" representation of an IPTW estimator in longitudinal MSM models with time-varying covariates.

Example (Continued): Further Robustness in Non-Ignorable Models. Consider model R0 in which model (3) for $\text{pr}(\Delta = 1|L)$ is replaced by logit $\pi(L; \rho_{\text{miss}}) = \eta(\mathbf{V}) + r(L)$, where $r(L)$ is a known function of L (such as $\alpha_0 Y$), $\eta(\mathbf{V})$ is an unknown function of $\mathbf{V}$ and $\rho_{\text{miss}} = \eta(\cdot)$. Model R0 is the discrete time (single occasion) version of model $A(\alpha_0)$ in our paper. Under model R0 μ is identified and equal to both $E[\Delta Y + (1-\Delta)E(Y \exp\{-r(L)\}|\Delta = 1, \mathbf{V})/E(\exp\{-r(L)\}|\Delta = 1, \mathbf{V})]$ and $E[\Delta Y/\text{pr}(\Delta = 1|L)]$. Furthermore, the semiparametric variance bound for μ is finite. However, estimators of μ that perform well for all laws allowed by the model do not exist due to the curse of dimensionality. To reduce dimensionality, we can consider either a parametric model $\eta(\mathbf{V}; \gamma)$ for $\eta(\mathbf{V})$ indexed by a q-dimensional parameter γ, or a parametric model $u(\mathbf{V}; \zeta)$ for the ratio $u(\mathbf{V}) = E[Y \exp\{-r(L)\}|\Delta = 1, \mathbf{V}]/E[\exp\{-r(L)\}|\Delta = 1, \mathbf{V}]$ indexed by an s-dimensional parameter ζ. The former determines a parametric non-response model R1 with $\rho_{\text{miss}} \equiv \gamma$. The latter is a semiparametric model for $f(Y|\Delta = 1, L)$ which we call a semiparametric pattern-mixture model. Now $\hat{\mu}(\hat{b})$ is a CAN estimator of μ when $\eta(\mathbf{V}; \gamma)$ is correctly specified. Furthermore, $\tilde{\mu}(g) = n^{-1} \sum_i \Delta_i Y_i + (1 - \Delta_i) u(\mathbf{V}_i; \hat{\zeta}(g))$ is a CAN estimator of μ when $u(\mathbf{V}; \zeta)$ is correctly specified, where $\hat{\zeta}(g)$ is a CAN estimator of ζ solving $\sum_i \Delta_i g(\mathbf{V}_i) \exp\{-r(L_i)\}\{Y_i - u(\mathbf{V}_i; \zeta)\} = 0$ with $g(\mathbf{V})$ any user-supplied s-dimensional function. Since we cannot be certain that these models are correctly specified, the best that can be hoped for is to find a single robust estimator that is CAN whenever either model $\eta(\mathbf{V}; \gamma)$ or model $u(\mathbf{V}; \zeta)$ is correct. Interestingly, it can be shown that the estimator $\hat{\mu}(\hat{b}_{\text{robust}})$ satisfies this hope, where $\hat{b}_{\text{robust}}(\mathbf{V}_i; \mu)$ is any user-supplied $1 + q$ dimensional function with first component equal to $u(\mathbf{V}_i; \hat{\zeta}(g)) - \mu$.

When $r(L)$ is identically 0 so that the data are CAR, $u(\mathbf{V}) = E[Y|\mathbf{V}]$ and the function $\hat{b}_{\text{eff}}$ of the earlier CAR example equals $\hat{b}_{\text{robust}}$. Thus, when $r(L) \equiv 0$, $\hat{\mu}(\hat{b}_{\text{robust}})$ is locally efficient at the submodel $u(\mathbf{V}; \zeta)$ in the semiparametric model characterized by the model logit$\pi(L; \rho_{\text{mis}}) = \eta(\mathbf{V}; \gamma) + r(L)$ and $r(L)$ known. When $r(L)$ is not identically 0, $\hat{\mu}(\hat{b}_{\text{robust}})$ is not locally efficient in this model; any locally efficient estimator will be inconsistent if the model $\eta(\mathbf{V}; \gamma)$ is misspecified even when the model $u(\mathbf{V}; \zeta)$ is correct.

3.2.4 Robustness of $\hat{\mu}(b)$ When the Data are Missing Completely at Random. When R is independent of L, we say the data are MCAR. Because when the true data-generating mechanism is MCAR, the complete-case estimator $\hat{\mu}_{\text{cc}}$ solving $0 = \sum_i \Delta_i m(L_i, \mu)$ is CAN, it seems important to require that any estimator used to adjust for potential selection bias due to measured or unmeasured factors be guaranteed to be CAN if in fact the data are MCAR. Otherwise, our estimator could fail where the simple complete-case analysis would succeed. In both ignorable and non-ignorable selection models the estimators $\hat{\mu}(\hat{b})$, in contrast to $\hat{\mu}_{\text{MLE}}$, are CAN under MCAR, provided that the model $f(r|l; \rho_{\text{mis}})$ includes all MCAR mechanisms.

Example (Continued). The nonignorable model (3) includes all MCAR mechanisms. To see this, set $\gamma_1 = \alpha = 0$ and vary γ_0. Hence $\hat{\mu}(\hat{b})$ is CAN under MCAR.

3.2.5 Robustness of $\hat{\mu}_{\text{loc,eff}}$ When the Positivity Assumption (1) Fails. Assumption (1) above is equivalent to assuming that (a) there are subjects with complete data and that (b) the support of L for subjects with coarsened data $(\Delta = 0)$ is included in that of subjects with complete data $(\Delta = 1)$. Of course, neither of these assumptions is guaranteed to hold. For example, (a) fails for the semiparametric CAR current status data model studied by van der Laan and Robins (1998) in which we observe both whether an underlying failure time variable $\mathcal{T}$ exceeds a random monitoring time Q and the history of a covariate process $\bar{\mathbf{V}}(t)$ until Q. Furthermore, (b) may fail in the nonselection mean model (16) considered in Section 7.3.2 of our article. Nonetheless,

as shown in these papers, both of these models admit robust RCAN estimators of the mean μ of $\mathcal{T}$ and Y. The point is that semiparametric estimators can be useful regardless of whether they allow the AIPCW representation of (2).

In the case of selection models in which (a) is true but (b) is false, it is standard practice to estimate μ by the MLE $\hat{\mu}_{\mathrm{MLE}}$ in a fully parametric model R3 rather than by an AIPCW estimators $\hat{\mu}(\hat{b})$ in a parametric nonresponse model R1 because, under model R1, $h(O;\mu,\rho_{\mathrm{mis}};b)$ may no longer have mean 0 and thus $\hat{\mu}(b)$ can be inconsistent. Surprisingly, even in this setting, in CAR models $\hat{\mu}_{\mathrm{loc,eff}} = \hat{\mu}(\hat{b}_{\mathrm{eff}})$ can be more robust to model misspecification than $\hat{\mu}_{\mathrm{MLE}}$, as illustrated by the following example.

Example (Continued). Suppose that $\mathrm{pr}[\Delta = 1|Y,\mathbf{V}] = \pi(\mathbf{V})I(\mathbf{V}\in\mathcal{K})$ where $\mathcal{K}$ is a known set and we specify a parametric CAR nonresponse model $\pi(\mathbf{V})\in\{\pi(\mathbf{V};\rho_{\mathrm{mis}})\}$. Thus the data are CAR, but the probability of observing Y is 0 for $\mathbf{V}\notin\mathcal{K}$, so the mean μ of Y is not even identified without strong assumptions. In this setting it is enormously difficult to specify either a correct parametric model $f(l;\rho_{\mathrm{ful}})$ for the joint law of $L=(Y,\mathbf{V})$ or a correct model $e(\mathbf{V};\omega)$ for $E[Y|\mathbf{V}]$, because of the need to extrapolate to the complement $\mathcal{K}^{\mathrm{c}}$ of $\mathcal{K}$. But again the model $e(\mathbf{V};\omega)$ is less restrictive than the model $f(l;\rho_{\mathrm{ful}})$. It can be shown that the estimator $\hat{\mu}(\hat{b}_{\mathrm{eff}})$ of Section 3.2.3 remains a CAN estimator of μ if ω is identified and $e(\mathbf{V};\omega)$ is correct regardless of whether the model $\pi(\mathbf{V};\rho_{\mathrm{mis}})$ is misspecified. In contrast, the MLE $\hat{\mu}_{\mathrm{MLE}}$ under $f(l;\rho_{\mathrm{ful}})$ generally will be consistent only if μ is identified and the model $f(l;\rho_{\mathrm{ful}})$ is correct. When model $e(\mathbf{V};\omega)$ is correct, the CAN estimator $n^{-1}\sum_i e(\mathbf{V}_i;\hat{\omega})$ is more efficient than $\hat{\mu}(\hat{b}_{\mathrm{eff}})$ except, when as in the example of Section 3.2.3, they are equal.

3.2.6 Lack of Robustness of Parametric Multiple Imputation. In parametric MI, one imputes missing values based on assuming a fully parametric model to create m completed datasets. The estimator $\hat{\mu}_{\mathrm{MI}}$ of the mean μ is then the mean of the m dataset-specific sample averages of Y (Rubin 1987). It follows that the qualitative robustness problems of $\hat{\mu}_{\mathrm{MLE}}$ are inherited by $\hat{\mu}_{\mathrm{MI}}$. However, when the amount of missing data is quite small, $\hat{\mu}_{\mathrm{MI}}$ generally will be much less biased than $\hat{\mu}_{\mathrm{MLE}}$ under misspecification of $f(l;\rho_{\mathrm{ful}})$, because $\hat{\mu}_{\mathrm{MI}}$, in contrast to $\hat{\mu}_{\mathrm{MLE}}$, does not replace a responder's observed Y with its predicted value under the model.

3.2.7 $\hat{\rho}_{\mathrm{MLE}}$ in a Parametric Model is an Augmented Inverse Probability of Censoring Weighted Estimator. Proposition A of Rotnitzky and Robins (1997) implies that in a parametric model R3, any RCAN estimator of ρ is asymptotically equivalent to an AIPCW estimator $\hat{\rho}\equiv\hat{\rho}(q,b)$ solving $\sum_i h(O_i;\rho;q,b)=0$, where $h(O;\rho;q,b)=\Delta\pi^{-1}(L;\rho)\{q(L;\rho)-E_\rho[(1-\Delta)b(O;\rho)|L]\}+(1-\Delta)b(O;\rho)$; $q(L;\rho)$ and $b(O;\rho)$ are vectors of the dimension of ρ and $q(L;\rho)$ is an unbiased estimating function; that is, $E_\rho[q(L;\rho)]=0$. The MLE $\hat{\rho}_{\mathrm{MLE}}$ is algebraically identical to the optimal AIPCW estimator $\hat{\rho}(q_{\mathrm{opt}},b_{\mathrm{opt}})$, where $b_{\mathrm{opt}}(O;\rho)=E_\rho[S^F(\rho)|O]$, $S^F(\rho)=\partial\log f(R,L;\rho)/\partial\rho$ is the full data score for ρ, and $q_{\mathrm{opt}}(L;\rho)=\pi(L;\rho)S^F(\rho)+E_\rho[(1-\Delta)b_{\mathrm{opt}}(O;\rho)|L]$. This can be proved by noting that in any missing-data model, the observed data score $S(\rho)$ equals $E_\rho[S^F(\rho)|O]$ and checking that $S(\rho)=h(O;\rho;q_{\mathrm{opt}},b_{\mathrm{opt}})$.

3.2.8 Semiparametric Complete-Data Models. In this section we generalize models R0 and R1 to allow for semiparametric complete-data models $f(l;\rho_{\mathrm{ful}})$, where $\rho_{\mathrm{ful}}=(\mu,\theta)$, μ is a finite-(say p) dimensional parameter of interest and θ is an infinite-dimensional nuisance parameter. We restrict attention to models $f(l;\rho_{\mathrm{ful}})$ in which in the absence of missing data, all RCAN estimators of μ are, up to asymptotic equivalence, equal to solutions $\hat{\mu}(m)$ to $\sum_i m(L_i;\mu)=0$, where $m(\cdot;\cdot)$ is a member of the set $\mathcal{M}$ of all p-dimensional unbiased estimating functions for μ; that is, $\mathcal{M}=\{m(L;\mu)\colon E_{\mu,\theta}[m(L;\mu)]=0 \text{ for all } \theta,\mu\}$. These models include the nonparametric complete-data models studied earlier as special cases. Robins et al. (1999) and Rotnitzky and Robins (1997) considered more general models.

Example (Continued). Consider the semiparametric regression model for $L=(Y,\mathbf{V}')'$ with $\mathbf{V}'=(Z',\mathbf{W}')$ characterized by the restriction $E[Y|Z]=g(Z;\mu_0)$, where $g(Z;\mu)$ is a known function; for example, $g(Z;\mu)=\mu'Z$. In this model, θ indexes all joint laws for $\varepsilon(\mu)\equiv Y-g(Z;\mu)$ and $\mathbf{V}$ that are restricted only by the condition that $E_{\mu,\theta}[\varepsilon(\mu)|Z]=0$ for all μ and θ, and $\mathcal{M}=\{m(L;\mu)=m(Z)[Y-g(Z;\mu)]\colon m(Z) \text{ a } p\text{-dimensional function}\}$. Any RCAN estimator of $\psi=(\mu,\rho_{\mathrm{mis}})$ in models R0 or R1 is asymptotically equivalent to an AIPCW estimator $\hat{\psi}\equiv\hat{\psi}(\hat{b})=(\hat{\mu}'(\hat{b}),\hat{\rho}'_{\mathrm{mis}}(\hat{b}))'$ solving (2), where now $\psi,\hat{b},d$, and h are $(p+q)$ dimensional and $d(L;\mu)$ is any $(p+q)$-dimensional estimating function for μ.

3.2.9 Sketch of Proof of $\hat{\mu}_{\mathrm{loc,eff}}$ in the Coarsened at Random Selection Models of Section 3.2.3. Define the operator $\mathbf{m}_\rho$ and its inverse $\mathbf{m}_\rho^{-1}$ as follows. Given a law $f(r,l;\rho)=f(r|l;\rho_{\mathrm{mis}})f(l;\rho_{\mathrm{ful}})$, for any $D=d(L)$, define $\mathbf{m}_\rho\{D\}=E_{\rho_{\mathrm{mis}}}[E_\rho\{D|O\}|L]$. When (1) holds, $\mathbf{m}_\rho$ is injective and the inverse operator $\mathbf{m}_\rho^{-1}$ exists. An important fact about distributions satisfying CAR, which is the key to our proof, is that conditional expectation of $d(L)$ given O depends only on the marginal law of L. That is, if $f(r|l;\rho_{\mathrm{mis}})$ satisfies CAR, then $E_\rho\{d(L)|O\}=E_{\rho_{\mathrm{ful}}}\{d(L)|O\}$.

Example (Continued). Rotnitzky and Robins (1997) showed that $\mathbf{m}_\rho^{-1}\{D\}=\pi^{-1}(L;\rho_{\mathrm{mis}})D+\{1-\pi^{-1}(L;\rho_{\mathrm{mis}})\}E^*_\rho\{D|\mathbf{V}\}$.

We restrict attention to the setup and models of Section 3.2.8. We consider a semiparametric selection model characterized by a parametric CAR nonresponse model $f(r|l;\rho_{\mathrm{mis}})$ and a semiparametric complete-data model $f(l;\rho_{\mathrm{ful}})$ with p-dimensional Euclidean parameter μ. Given a p-dimensional function $m\in\mathcal{M}$ and a parametric submodel of the model $f(l;\rho_{\mathrm{ful}})$, define the random function $\hat{b}^m(O;\mu)=E_{\hat{\rho}_{\mathrm{ful},r\mathrm{MLE}(\mu)}}[\mathbf{m}^{-1}_{\hat{\rho}_{\mathrm{ful},\mathrm{MLE}(\mu)},\hat{\rho}_{\mathrm{mis}}}\{m(L,\mu)\}|O]$,

where $\hat{\rho}_{\text{ful},r\text{MLE}}(\mu)$ is the MLE of ρ_{ful} in the parametric submodel with μ held fixed and $\hat{\rho}_{\text{mis}}$ is the MLE of ρ_{mis}. Let $\hat{\mu}^m$ solve $\sum_i \hat{b}^m(O_i;\mu) = 0$. Note that $\hat{\mu}^m$ is equal to the solution $\hat{\mu}(\hat{b}^m)$ to (2) when $\mu, \hat{\rho}_{\text{mis}}, m$, and $\hat{b}^m$ are substituted for $\psi, \rho_{\text{mis}}, d$, and $\hat{b}$. Robins et al. (1994) proved that $\hat{\mu}^{\hat{m}_{\text{eff}}}$ is a locally efficient estimator $\hat{\mu}_{\text{loc,eff}}$ of μ at the parametric complete-data submodel, where $\hat{m}_{\text{eff}}$ converges to a certain $m_{\text{eff}} \in \mathcal{M}$. Hence, to prove robustness of $\hat{\mu}_{\text{loc,eff}}$ in the sense described in Section 3.2.3, it suffices to prove the following.

Theorem. For $\hat{m}$ converging to an $m \in \mathcal{M}, \hat{\mu}^{\hat{m}}$ is a CAN estimator of μ if either (a) the parametric CAR model $f(r|l;\rho_{\text{mis}})$ and the semiparametric model $f(l;\rho_{\text{ful}})$ are both correct or (b) the parametric submodel of model $f(l;\rho_{\text{ful}})$ is correct and the true data-generating process satisfies CAR.

Proof Sketch. Let $\rho^\top = (\rho_{\text{ful}}^\top, \rho_{\text{mis}}^\top)$ denote the true CAR law generating the data, and let ρ_{mis}^* and ρ_{ful}^* be the probability limits of $\hat{\rho}_{\text{mis}}$ and $\hat{\rho}_{\text{ful},r\text{MLE}}(\mu^\top)$ and set $\rho^* = (\rho_{\text{ful}}^*, \rho_{\text{mis}}^*)$. Note that $\rho_{\text{mis}}^* = \rho_{\text{mis}}^\top$ if model $f(r|l;\rho_{\text{mis}})$ is correct, and $\rho_{\text{ful}}^* = \rho_{\text{ful}}^\top$, if the parametric submodel is correct. Thus, under regularity conditions, the theorem will be true if we can show for all $m \in \mathcal{M}, U \equiv E_{\rho^\top}[E_{\rho_{\text{ful}}^*}[\mathbf{m}_{\rho^*}^{-1}\{m(L,\mu^\top)\}|O]] = 0$ when either $\rho_{\text{mis}}^* = \rho_{\text{mis}}^\top$ or $\rho_{\text{ful}}^* = \rho_{\text{ful}}^\top$. If $\rho_{\text{mis}}^* = \rho_{\text{mis}}^\top$, then $U = E_{\rho_{\text{ful}}^\top}[\mathbf{m}_{\rho^*}\mathbf{m}_{\rho^*}^{-1}\{m(L,\mu^\top)\}] = E_{\rho_{\text{ful}}^\top}[m(L,\mu^\top)] = 0$ by $m \in \mathcal{M}$. Next, suppose that $\rho_{\text{ful}}^* = \rho_{\text{ful}}^\top$. Because, under CAR, $E_{\rho_{\text{ful}}^\top,\rho_{\text{mis}}^\top}[E_{\rho_{\text{ful}}^\top}\{d(L)|O\}] = E_{\rho_{\text{ful}}^\top,\rho_{\text{mis}}^*}[E_{\rho_{\text{ful}}^\top}\{d(L)|O\}] = E_{\rho_{\text{ful}}^\top}[d(L)]$, we obtain $U = E_{\rho_{\text{ful}}^\top,\rho_{\text{mis}}^*}[E_{\rho_{\text{ful}}^*}[\mathbf{m}_{\rho^*}^{-1}\{m(L,\mu^\top)\}|O]] = E_{\rho_{\text{ful}}^\top,\rho_{\text{mis}}^*}[\mathbf{m}_{\rho^*}\mathbf{m}_{\rho^*}^{-1}\{m(L,\mu^\top)\}] = E_{\rho_{\text{ful}}^\top,\rho_{\text{mis}}^*}[m(L,\mu^\top)] = 0$.

Remark. The results in this section remain true, except for minor notational changes, if we substitute a semiparametric nonresponse model R0 for the parametric nonresponse model R1. In models in which the positivity assumption (1) does not hold, $\mathbf{m}_\rho$ is not injective. Nonetheless, under regularity conditions, the efficient influence function for μ still has the form $E_\rho\{\mathbf{m}_\rho^{-1}[m_{\text{eff}}(L,\mu)]|O\}$, except that $\mathbf{m}_\rho^{-1}$ is now a generalized inverse (van der Vaart 1991); in that case the results in this section remain true, except that $\hat{\mu}^m$ may no longer have a representation as an AIPCW estimator $\hat{\mu}(\hat{b}^m)$. Finally, Robins et al. (1999) and Rotnitzky and Robins (1997) showed how to calculate $\hat{m}_{\text{eff}}(L,\mu) = \hat{m}_{\text{eff}}(Z)[Y - g(Z;\mu)]$ and $\hat{\psi}_{\text{loc,eff}}$ in the regression model $E[Y|Z] = g(Z;\mu_0)$ of Section 3.2.8 in both CAR and non-CAR models for various data structures treated in our article and in our example.

3.3 Ignorability and Nonignorability in Model (15) With $\alpha_0 = 0$

Consider the monotone missing-data structure of Section 3.1. The full data $L = (\bar{\mathbf{V}}(\mathcal{T}), \mathcal{T})$ are censored by Q, resulting in observed data $O = (X = \min(\mathcal{T}, Q), \Delta = I(X = \mathcal{T}), \bar{\mathbf{V}}(X))$. A model $\lambda_Q(t|\bar{\mathbf{V}}(t);\omega)$ for the cause-specific hazard of censoring given the observed past $\bar{\mathbf{V}}(t)$ is said to be *ignorable for inference about* μ if the observed data likelihood $\mathcal{L}(O;\rho)$ with $\rho = (\mu, \xi, \omega)$ factorizes as $\mathcal{L}(O;\rho) = \mathcal{L}_1(O;\mu,\xi)\mathcal{L}_2(O;\omega)$ and (μ,ξ) and ω are "distinct," because then the likelihood factor $\mathcal{L}_2(O;\omega)$ can be "ignored" for inference about μ. Two parameter vectors, ρ_1 and ρ_2, are distinct for frequentist inference if they are variation independent and for Bayesian inference if they have independent priors (Rubin 1976). Rubin (1976) noted that CAR selection models with $\rho_{\text{ful}} = (\mu,\xi)$ and $\rho_{\text{mis}} = \omega$ are ignorable. (Robins and Ritov (1997) showed, however, that, in this setting, if ξ is infinite dimensional, then the factor $\mathcal{L}_2(O;\omega)$ often cannot be ignored for frequentist inference about μ. Nonetheless, we shall continue to employ the usual nomenclature and refer to the model $\lambda_Q(t|\bar{\mathbf{V}}(t);\omega)$ as ignorable even when ξ is infinite dimensional.)

Interestingly, (15) with $\alpha_0 = 0$, which we call model $A^*(0)$, provides an example of an *ignorable non-CAR model*. To illustrate this point in a simple setting, we use the discrete time model described earlier in Section 2.5.4. To do so, we identify our $\mathbf{V}(1), \mathbf{V}(2)$, and $Y \equiv \mathbf{V}(3)$ with LR's Y_1, Y_2, and Y_3 and set $\mathcal{T}$ equal to 3 with probability 1. To simplify the exposition, we assume that Y_3 is dichotomous. Under model $A^*(0)$, the discrete hazard $\lambda_Q(t|Y_1, Y_t, Y_3;\omega) \equiv \text{pr}[Q = t|Y_1, Y_t, Y_3, Q \geq t;\omega]$ is assumed to not depend on Y_3 and thus equals $\lambda_Q(t|Y_1, Y_t;\omega)$ for $t = 1, 2$. Hence the individual likelihood contribution $\mathcal{L}(Q, L;\rho)$ based on data $(Q, L = (Y_1, Y_2, Y_3))$ can be factorized as $\mathcal{L}(Q, L;\rho) = \mathcal{L}^{\text{ful}}(Q, L;\mu,\xi)\mathcal{L}^{\text{mis}}(O;\omega)$, where $\mathcal{L}^{\text{ful}}(Q, L;\mu,\xi) = \mathcal{L}_1^{\text{ful}}(Q, L;\mu,\phi)\mathcal{L}_2^{\text{ful}}(Q, L;\nu)$, with $\mathcal{L}_1^{\text{ful}}(Q, L;\mu,\phi) = f(Y_3;\mu)f(Y_1|Y_3;\phi_1)f(Y_2|Y_1, Y_3, Q \neq 1;\phi_2)^{I(Q\neq 1)}$, $\mathcal{L}_2^{\text{ful}}(Q, L;\nu) = \{f(Y_2|Y_1, Y_3, Q = 1;\nu)\}^{I(Q=1)}$, $\mathcal{L}^{\text{mis}}(O;\omega) = \lambda_Q(1|Y_1;\omega)^{I(Q=1)}[\{1-\lambda_Q(1|Y_1;\omega)\}\lambda_Q(2|Y_1, Y_2;\omega)^{I(Q=2)}\{1 - \lambda_Q(2|Y_1, Y_2;\omega)\}^{I(Q\neq 2)}]^{I(Q\neq 1)}$, $\xi = (\phi,\nu)$, and $\phi = (\phi_1,\phi_2)$. It follows that the observed data likelihood is $\mathcal{L}(O;\rho) = \mathcal{L}_1(O;\mu,\phi)\mathcal{L}_2(O;\omega)$, where $\mathcal{L}_1(O;\mu,\phi) = [\mathcal{L}_1^{\text{full}}(Q, L;\mu,\xi)]^{I(Q=3)}[\sum_{Y_3=0}^1 \mathcal{L}_1^{\text{full}}(Q, L;\mu,\xi)]^{\mathcal{I}(Q\leq 2)}$ and $\mathcal{L}_2(O;\omega) = \mathcal{L}^{\text{mis}}(O;\omega)$. Examination of these likelihoods reveals that: (a) $\mathcal{L}(O;\rho)$ does not depend on ν, so ν is not identified from the observed data O; (b) the "implausible" conditional independence assumption $I(Q = 1) \coprod Y_2|Y_1, Y_3$ implied by model (1) of Section 2.5.4 is equivalent to $\nu = \phi_2$ and thus the data have nothing to say about its validity; (c) when (μ,ϕ) and ω are distinct, the model $\lambda_Q(t|Y_1, Y_t;\omega)$ is ignorable; (d) $A^*(0)$ is a CAR model [and identical to the model of equation (1) of our paper with $\alpha_0 = 0$] if and only if $\nu = \phi_2$. Items (c) and (d) imply that a submodel of $A^*(0)$ in which (μ,ϕ_1,ϕ_2) and ω are distinct and $\nu = \phi_2 - c$ for some known nonzero constant c is non-CAR but ignorable. Furthermore, if we reparameterize the likelihood $\mathcal{L}(Q, L;\rho)$ as $f(Q|L;\rho_{\text{mis}})f(L;\rho_{\text{ful}})$, then it can be shown that ρ_{ful} and ρ_{mis} are not distinct in this submodel, which exemplifies the fact that an ignorable non-CAR model cannot be a selection model. On the other hand, suppose that we consider a submodel of model $A^*(0)$ in which ρ_{mis} and ρ_{ful} are distinct. Then if $\nu \neq \phi_2$, the model is non-CAR. But a non-CAR selection model is nonignorable for inference on μ. This is so even though the observed data likelihood still factorizes into a (μ,ϕ) part and a ω part, because now ω and (μ,ϕ) are not distinct. In summary, a

submodel of model $A^*(0)$ can be an ignorable CAR selection model, an ignorable non-CAR nonselection model, or a nonignorable non-CAR selection model, depending on whether or not we set ν equal to ϕ_2, and whether we assume that the pair ω and (μ, ϕ) versus the pair ρ_{mis} and ρ_{ful} is distinct.

3.4 Follow-Up of Nonrespondents

Suppose that in our AIDS example, the only censoring had been due to loss to follow-up. Further suppose that to diminish bias due to nonignorable drop-out, the investigators make up to K attempts (say, house visits) to contact the nonrespondents and measure their outcome Y. (These contacts are made soon after the end of follow-up time T at which the CD4 count Y was to be measured, so temporal trends in CD4 can be ignored.) In this setting the observed data are now $O = (C, Q, \bar{\mathbf{V}}(Q), \Delta, \Delta Y)$, where $C \equiv 0$ if a subject is a completer $(Q = T); C \equiv k$ if Y is measured on the kth contact attempt for $k = 1, 2, \ldots, K-1$, and $C \equiv K$ otherwise; and Δ is the indicator that Y is observed. Redefine $L \equiv (C, Q, \bar{\mathbf{V}}(Q), Y)$ and $\pi(L) = \text{pr}[\Delta = 1|L]$. Consider the semiparametric model R1 characterized by the model for $\pi(L)$,

$$\pi(L; \rho_{\text{mis}}) = expit[\kappa(Q, \bar{\mathbf{V}}(Q); \rho_{\text{mis}}) + r(L, \beta_0)]\}^{I(C=K)}, \quad (4)$$

where $expit(u) \equiv \{1 + \exp(-u)\}^{-1}, \kappa(Q, \bar{\mathbf{V}}(Q); \rho_{\text{mis}})$ is a known function, ρ_{mis} is an unknown finite-dimensional parameter, $r(L; \beta)$ is a known function satisfying $r(L; 0) = 0$, and we regard the nonignorable selection bias parameter β_0 as known but vary it in a sensitivity analysis. Note that, as required, $\pi(L; \rho_{\text{mis}}) = 1$ if $C \neq K$, because, by definition, $\Delta = \pi(L) = 1$ if $C \neq K$. Then, by proposition A1 of Rotnitzky and Robins (1997), up to asymptotic equivalence, all RCAN estimators of $\psi = (\mu, \rho'_{\text{mis}})'$ under (4) based on the data O can be obtained by solving $\sum_i h(O_i; \mu, \rho_{\text{mis}}; \hat{b}) = 0$ of (2) for some $\hat{b}(O; \psi)$. Note that we no longer need to correctly specify either model (13) or (15) of our article to obtain RCAN estimators of the mean μ of Y.

When it is too costly to attempt to contact all drop-outs, the multistage monotone random sampling design (hereafter, ms design) studied by Rotnitzky and Robins (1995) is often used. In this design, contact attempts are made on each nonrespondent until either K attempts have been made, Y is recorded, or the outcome of a known random sampling mechanism specifies that the contact attempts are to be discontinued. The theory of Rotnitzky and Robins (1997) can be used to obtain estimators of ψ as follows. Let S denote the (possibly unobserved) first occasion $k, k \in \{1, 2, \ldots, K+1\}$ at which the random sampling mechanism would specify that no contact be attempted, so S can "censor" C. Let $X_{\text{ms}} = \min(C, S)$ and $\Delta_{\text{ms}} = I(S > C)$. Under the ms design, the observed data are $O_{\text{ms}} = (X_{\text{ms}}, \Delta_{\text{ms}}, Q, \bar{\mathbf{V}}(Q), \Delta_{\text{ms}}\Delta, \Delta_{\text{ms}}\Delta Y)$. Now regarding $O = (C, Q, \bar{\mathbf{V}}(Q), \Delta, \Delta Y)$ as the "full" data that would be observed if all subjects had $\Delta_{\text{ms}} = 1$ (as was the case under the design analyzed in the preceding paragraph), the conditional probability $\text{pr}[\Delta_{\text{ms}} = 1|O]$ of observing the "full" data O under the ms design is $\Pi_{\text{ms}}(C)$, where $\Pi_{\text{ms}}(j) = \prod_{k=0}^{j}\{1 - \lambda_{\text{ms},k}(Q, \bar{\mathbf{V}}(Q))\}, \lambda_{\text{ms},0}(Q, \bar{\mathbf{V}}(Q)) \equiv 0$, and the random sampling probabilities $\lambda_{\text{ms},k}(Q, \bar{\mathbf{V}}(Q)) \equiv \text{pr}[S = k|X_{\text{ms}} \geq k, O] = \text{pr}[S = k|X_{\text{ms}} \geq k, Q, \bar{\mathbf{V}}(Q)]$ are known by design for $k = 1, \ldots, K$. Hence, by proposition A1 of Rotnitzky and Robins (1997), up to asymptotic equivalence, all RCAN estimators of ψ in the model characterized by (4) and data O_{ms} are given by $\hat{\psi}_{\text{ms}}(\hat{b}, \hat{b}_{\text{ms}})$ solving $0 = \sum_i h_{\text{ms}}(O_{\text{ms},i}; \psi; \hat{b}, \hat{b}_{\text{ms}})$, where $h_{\text{ms}}(O_{\text{ms}}; \psi; \hat{b}, \hat{b}_{\text{ms}}) = \Delta_{\text{ms}}\Pi_{\text{ms}}^{-1}(C)\{h(O; \mu, \rho_{\text{mis}}; \hat{b}) - E[(1 - \Delta_{\text{ms}})\hat{b}_{\text{ms}}(O_{\text{ms}}; \psi)|O]\} + (1 - \Delta_{\text{ms}})\hat{b}_{\text{ms}}(O_{\text{ms}}; \psi)$. This result assumes that each subject is sampled independently. Robins et al. (1994, sec. 6.4) provided an extension appropriate for nonindependent sampling designs.

It follows from proposition 8.2 of Robins et al. (1994) that for any choice of $\hat{b}(O; \psi)$ with probability limit $b(O; \psi)$, the locally optimal choice $\hat{b}^{\hat{b}}_{\text{ms}}$ of $\hat{b}_{\text{ms}}$ at a parametric model $f(L; \rho_{\text{ful}})$ is $\hat{b}^{\hat{b}}_{\text{ms}}(O_{\text{ms}}; \psi) = \Phi\{h(O; \mu, \rho_{\text{mis}}; \hat{b})\}$, where $\Phi(Z) = H^Z(X_{\text{ms}})/\Pi_{\text{ms}}(X_{\text{ms}}) - \sum_{j=0}^{X_{\text{ms}}} \lambda_{\text{ms},j}(Q, \bar{\mathbf{V}}(Q))H^Z(j)/\Pi_{\text{ms}}(j), H^Z(j) = E_{\hat{\rho}_{\text{ful},r\text{MLE}(\psi)}}[Z|Q, \bar{\mathbf{V}}(Q), C > j - 1]$ and $\hat{\rho}_{\text{ful},r\text{MLE}}(\psi)$ is the MLE of ρ_{ful} with ψ fixed. It then follows from Rotnitzky and Robins (1997) that a locally efficient AIPCW estimator $\hat{\psi}_{\text{loc,eff}}$ is obtained by choosing $\hat{b}$ to be $\hat{b}_{\text{eff}} = \hat{b}_{\text{eff}}(L; \psi) = I(C = K)E_{\hat{\rho}_{\text{ful},r\text{MLE}(\psi)}}[(\pi^{-1}(L; \rho_{\text{mis}}) - 1)H(\psi)|C = K, Q, \bar{\mathbf{V}}(Q)]/E_{\hat{\rho}_{\text{ful},r\text{MLE}(\psi)}}[(\pi^{-1}(L; \rho_{\text{mis}}) - 1)\Pi_{\text{ms}}(C)|C = K, Q, \bar{\mathbf{V}}(Q)]$ with $H(\psi) \equiv (Y - \mu, \pi(L; \rho_{\text{mis}})\partial\kappa(Q, \bar{\mathbf{V}}(Q); \rho_{\text{mis}})/\partial\rho'_{\text{mis}})'$. Putting all of the above together, we obtain that $\psi_{\text{loc,eff}}$ solves

$$0 = \sum_i \Delta_{\text{ms},i}\Delta_i \left\{\Pi_{\text{ms};i}(C_i)\pi(L_i; \rho_{\text{mis}})^{I(C_i=K)}\right\}^{-1} H_{1i}(\psi) + \sum_i \Delta_{\text{ms},i}\{\Delta_i\pi(L_i; \rho_{\text{mis}})^{-1} - 1\}\hat{b}_{\text{eff}}(L_i; \psi) + \Phi_i(H_{1i}(\psi)),$$

where $H_{1i}(\psi) = (Y_i - \mu, \mathbf{0}')'$ and $\mathbf{0}$ is a column vector of the dimension of ρ_{mis}.

Suppose that there are no persistent nonresponders ($\Delta = 1$ with probability), so all missing data on Y are by design. Then the foregoing procedure produces a locally efficient estimator when we take $\hat{b}_{\text{eff}}$ equal to 0. In this setting, several authors have proposed using the nonparametric maximum likelihood estimator (NPMLE) of the mean μ based on either the data O or on a subset of the data. However, due to the curse of dimensionality, the NPMLE, in contrast to the locally efficient AIPCW estimator, can neither recover the information about μ contained in the data $(Q, \bar{\mathbf{V}}(Q))$ nor allow for the known selection probabilities $\lambda_{\text{ms},k}$ to depend in a complex way on $(Q, \bar{\mathbf{V}}(Q))$ (Robins and Ritov 1997).

3.5 Additional Issues

3.5.1 Little and Rubin. LR reference an article by Rubin, Stern, and Vehovar (RSV) as support for their view that the MAR assumption may often be a reasonable approximation to reality thus obviating the need for a sensitivity analysis. However, we do not find that the RSV article provides meaningful support for LR's views. RSV

analyzed a poll with nonmonotone missing data whose purpose was to predict the result of the Slovenian plebiscite that occurred 3 weeks later. RSV offered two arguments. The first was based on designating one particular unsaturated nonignorable model as "obvious" and demonstrating that this model failed to fit the data. RSV then fit a saturated MAR model (which, by necessity, fits the data perfectly). From these facts, they concluded that their MAR assumption was reasonable, although they noted in passing that there would be other nonignorable models whose likelihood would attain that of the saturated MAR model, but designated these models as "not compelling." However, they provided no scientific rationale as to why their ill-fitting nonignorable model was more compelling than other possibly well-fitting nonignorable models. Indeed, in a reanalysis of this data, Molenberghs, Kenward, and Goetghebeur (1999) discovered other well-fitting models and argued that these appear no less a priori plausible than RSV's model. Further, a number of these well-fitting nonignorable models predicted vote counts that differed substantially from the count predicted by the saturated MAR model endorsed by RSV. Finally, Robins and Gill (1997) questioned the plausibility of RSV's saturated MAR model itself, showing that mechanisms that generate nonmonotone missing MAR data are quite special and in particular do not include the latent trait factor-analytic model mentioned by RSV.

RSV's second argument was that the results obtained based on the saturated MAR analysis, in contrast to those based on their nonignorable model, agreed well with the results of the actual plebiscite held 3 weeks later. We find this argument unconvincing, as it is essentially a meta-analytic one, but based on a single study. For even if missingness were MAR for the Slovenian survey, what does that tell us about other surveys, much less other substantive areas, such as ACTG trial 175? Further, why should one assume that in a changing political situation, the result that would have been obtained in the survey had there been no missing data (which, as stressed by RSV, is the inferential goal of any method to correct for missing data) would have agreed with that obtained 3 weeks later in the actual plebiscite? It is possible that the nonignorable model results were closer to the inferential goal.

3.5.2 Properties of Laird and Pauler's Estimator. Under LP's discrete time version of (16), drop-out can occur only at times $t = 8 + 12u, u = 0, 1, 2, 3$. For $u = 3, 2, 1, 0$, let $H(u)$ be recursively defined by $H(u) = E[H(u+1)|\bar{\mathbf{V}}(8+12u), Q > 8+12u] + \phi\lambda[8+12u|\bar{\mathbf{V}}(8+12u)]$ with $H(4) \equiv Y$. Then, under LP's model, $E[Y] = E[H(0)]$ and $E[Y|\bar{\mathbf{V}}(8+12u), Q > 8+12u] = E[H(u+1)|\bar{\mathbf{V}}(8+12u), Q > 8+12u]$. These identities motivate the following estimator of μ_0 under LP's suggested approach. One specifies regression models $m_t(\bar{\mathbf{V}}(t), \beta_t)$ for $E[Y|\bar{\mathbf{V}}(t), Q > t]$. Recursively define, for $u = 3, 2, 1, 0, \hat{H}(u) = m_{8+12u}(\bar{\mathbf{V}}(8+12u), \hat{\beta}_{8+12u}) + \phi\hat{\lambda}[8+12u|\bar{\mathbf{V}}(8+12u)]$, where $\hat{\lambda}[8+12u|\bar{\mathbf{V}}(8+12u)]$ is the empirical (nonparametric) estimator of the discrete hazard $\lambda[8+12u|\bar{\mathbf{V}}(8+12u)]$, and $\hat{\beta}_{8+12u}$ is obtained by possibly nonlinear regression of $\hat{H}(u+1)$ on $\bar{\mathbf{V}}(8+12u)$ among subjects with $Q > 8+12u$ with $\hat{H}(4) \equiv Y$. When $\mathbf{V}(u)$ has a continuous distribution then $\hat{\lambda}[8+12u|\bar{\mathbf{V}}(8+12u)]$ is either 0 or 1, that is $\hat{\lambda}[8+12u|\bar{\mathbf{V}}(8+12u)] = I\,(Q = 8+12u)$. Then $\mu_0 = E[Y]$ is estimated by $n^{-1}\sum_i \hat{H}_i(0)$. For consistency, the regression model $m_t(\bar{\mathbf{V}}(t), \beta_t)$ needs to be correct even under MCAR. When $\phi = 0$, this estimator is the ICE estimator discussed by Robins et al. 1995 and Robins 1998. The discrete time version of our estimator of μ_0 described in Section 7.3.2 is $n^{-1}\sum_i \Delta_i\{Y_i + \phi\sum_{u=0}^{3}\tilde{\lambda}[8+12u|\bar{\mathbf{V}}_i(8+12u)]\}/\prod_{u=0}^{3}\{1-\tilde{\lambda}[8+12u|\bar{\mathbf{V}}_i(8+12u)]\}$ where $\tilde{\lambda}[8+12u|\bar{\mathbf{V}}_i(8+12u)]$ is an estimate of the hazard based on a finite dimensional parametric model that leaves the baseline hazard unrestricted. In contrast, because the estimator of the mean based on LP's suggested approach is linear in the estimate of $\lambda[8+12u|\bar{\mathbf{V}}(8+12u)]$, it was possible to use a nonparametric estimator of the hazard without sacrificing consistency.

To see why the models $m_t(\bar{\mathbf{V}}(t), \beta_t)$ will often be mutually incompatible, suppose Y is dichotomous, $\mathbf{V}(8)$ is univariate and continuous and all other $\mathbf{V}(t)$ are dichotomous. If we choose $\phi = 0$ and each $m_t(\bar{\mathbf{V}}(t), \beta_t)$ to be a linear logistic function of the components of $\bar{\mathbf{V}}(t)$, then it is easy to show there is no joint distribution for which each component of β_t is non zero for $t \in \{8, 20, 32, 44\}$.

ADDITIONAL REFERENCES

Clayton, D., Spiegelhalter, D., Dunn, G., and Pickles, A. (1998), "Analysis of Longitudinal Binary Data From Multiphase Sampling," *Journal of the Royal Statistical Society*, Ser. B, 60, 71–87.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1996), "Coarsening at Random: Characterizations, Conjectures and Counterexamples," *Proceedings of the First Seattle Symposium on Survival Analysis*, pp. 255–294.

Molenberghs, G., Kenward, M., and Goetghebeur, E. (1999), "Sensitivity Analysis for Incomplete Contingency Tables," in press.

Robins, J. M., (1995), "An Analytic Method for Randomized Trials With Informative Censoring: Part I," *Lifetime Data Analysis*, 1, 241–254.

——— (1998), "Correction for Non-Compliance in Equivalence Trials," *Statistics in Medicine*, 17, 269–302.

——— (1999), "Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference," in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, eds. M. E. Halloran and D. Berry, New York: Springer-Verlag, pp. 95–134.

——— (2000), "Robust Estimation of Sequentially Ignorable Missing Data and Causal Inference Models," Proceedings of the 1999 American Statistical Association Joint Meetings, to appear.

Robins, J. M., and Gill, R. (1997), "Non-Response Models for the Analysis of Non-Monotone Ignorable Missing Data," *Statistics in Medicine*, 16, 39–56.

Robins, J. M., Greenland, S., and Hu, F. C. (1999). Rejoinder to Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. To appear in the Journal of the American Statistical Association.

Robins, J. M., and van der Laan, M. (to appear), Discussion of "On Profile Likelihood" by S. A. Murphy and A. W. van der Vaart, *Journal of the American Statistical Association*, .

Robins, J. M., and Wang, N. (1998), Discussion of D. Clayton et al., *Journal of the Royal Statistical Society*, Ser. B, 60, 91–93.

Rotnitzky, A., and Robins, J. M. (1995), "Semiparametric Regression With Follow-Up of Nonrespondents," unpublished mimeo.

van der Laan, M. J., and Robins, J. M. (1998), "Locally Efficient Estimation With Current Status Data and Time-Dependent Covariates," *Journal of the American Statistical Association*, 93, 693–701.

van der Vaart, A. W. (1991), "On Differentiable Functionals," *The Annals of Statistics*, 19, 178–204.