Statistical Science
2005, Vol. 20, No. 3, 261–301
DOI 10.1214/088342305000000151

Semiparametric Estimation of Treatment Effect in a Pretest–Posttest Study with Missing Data

Marie Davidian, Anastasios A. Tsiatis and Selene Leon

Abstract. The pretest–posttest study is commonplace in numerous applications. Typically, subjects are randomized to two treatments, and response is measured at baseline, prior to intervention with the randomized treatment (pretest), and at prespecified follow-up time (posttest). Interest focuses on the effect of treatments on the change between mean baseline and follow-up response. Missing posttest response for some subjects is routine, and disregarding missing cases can lead to invalid inference. Despite the popularity of this design, a consensus on an appropriate analysis when no data are missing, let alone for taking into account missing follow-up, does not exist. Under a semiparametric perspective on the pretest–posttest model, in which limited distributional assumptions on pretest or posttest response are made, we show how the theory of Robins, Rotnitzky and Zhao may be used to characterize a class of consistent treatment effect estimators and to identify the efficient estimator in the class. We then describe how the theoretical results translate into practice. The development not only shows how a unified framework for inference in this setting emerges from the Robins, Rotnitzky and Zhao theory, but also provides a review and demonstration of the key aspects of this theory in a familiar context. The results are also relevant to the problem of comparing two treatment means with adjustment for baseline covariates.

Key words and phrases: Analysis of covariance, covariate adjustment, influence function, inverse probability weighting, missing at random.

1. INTRODUCTION

1.1 Background and Motivation

The so-called pretest–posttest trial arises in a host of applications. Subjects are randomized to one of two interventions, denoted here by "control" and "treatment," and the response is recorded at baseline, prior to intervention (pretest response), and again after a prespecified follow-up period (posttest response). We use the terms "baseline/pretest" and "follow-up/posttest" interchangeably. The effect of interest is usually stated as "difference in change of (mean) response from baseline to follow-up between treatment and control."

For instance, in studies of HIV disease, a common objective is to determine whether the change in measures of immunologic status such as CD4 cell count from baseline to some subsequent time following initiation of antiretroviral therapy is different for different treatments. Depressed CD4 counts indicate impairment of the immune system, so larger, positive such changes are thought to reflect more effective treatment. To exemplify this situation, we consider data from 2139 patients from AIDS Clinical Trials Group (ACTG) protocol 175 (Hammer et al., 1996), a study that randomizes patients to four antiretrovi-

Marie Davidian and Anastasios A. Tsiatis are Professors, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (e-mail: davidian@stat.ncsu.edu; tsiatis@stat.ncsu.edu). Selene Leon is Senior Biostatistician, Novartis Pharmaceuticals, East Hanover, New Jersey 07936-1080, USA (e-mail: selene.leon@pharma.novartis.com).

ral regimens in equal proportions. The findings of ACTG 175 indicate that zidovudine (ZDV) monotherapy is inferior to the other three [ZDV+didanosine (ddI), ZDV+zalcitabine, ddI] therapies, which showed no differences on the basis of the primary study endpoint of progression to AIDS or death. Accordingly, we consider two groups: subjects who receive ZDV alone (control) and those who receive any of the other three therapies (treatment). As is routine in HIV clinical studies, measures such as CD4 count were collected on all participants periodically throughout, and interest also focused on secondary questions regarding changes in immunologic and virologic status. An important secondary endpoint was change in CD4 count from baseline to 96±5 weeks.

To formalize this situation, let Y_1 and Y_2 denote baseline and follow-up response (e.g., baseline and 96±5 week CD4 count) and let $Z = 0$ or 1 indicate assignment to control or treatment, respectively. Because, under proper randomization, pretest mean response should not differ by intervention, it is reasonable to assume that $E(Y_1|Z = 0) = E(Y_1|Z = 1) = E(Y_1) = \mu_1$. Letting $E(Y_2|Z) = \mu_2 + \beta Z$, the desired effect may then be expressed as

$$\begin{aligned}&\{E(Y_2|Z=1) - \mu_1\} - \{E(Y_2|Z=0) - \mu_1\}\\ &\quad = E(Y_2|Z=1) - E(Y_2|Z=0) = \beta\end{aligned} \tag{1}$$

and interest focuses on the parameter β. A number of ways to make inference on β have been proposed. Because the question is usually posed in terms of difference in change from baseline, analysis is often based on the "paired t test" estimator for β found by taking the difference of the sample averages of $(Y_2 - Y_1)$ in each group. The second expression in (1) involves only posttest treatment means, suggesting estimating β in the spirit of the two-sample t test by the difference of Y_2 sample means for each treatment (ignoring baseline responses altogether). However, if baseline response is correlated with change in response or posttest response itself, intuition suggests taking this into account. For continuous response, this has led many researchers to advocate the use of analysis of covariance (ANCOVA) techniques, in which one estimates β directly by fitting the linear model $E(Y_2|Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \beta Z$. A variation is to include an interaction term involving $Y_1 Z$; here, β is estimated as the coefficient of $(Z - \overline{Z})$ in the regression of $Y_2 - \overline{Y}_2$ on $Y_1 - \overline{Y}_1$, $Z - \overline{Z}$ and $(Y_1 - \overline{Y}_1)(Z - \overline{Z})$, where the overbars denote overall sample average. Singer and Andrade (1997) mentioned a "generalized estimating equation" (GEE) approach (see also Koch, Tangen, Jung and Amara, 1998), where $(Y_1, Y_2)^T$ is viewed as a multivariate response vector with mean $(\mu_1, \mu_2 + \beta Z)^T$ and standard GEE methods are used to make inference on β. Yang and Tsiatis (2001) and Leon, Tsiatis and Davidian (2003) provided further details on all of these methods. The two-sample t test approach implicitly assumes pre- and posttest responses are uncorrelated, which may be unrealistic, while the paired t test and ANCOVA evidently assume linear dependence of Y_1 and Y_2, which may not hold in practice; for example, Figure 1 shows baseline and follow-up CD4 counts in ACTG 175 and suggests a mild curvilinear relationship between them in each group. Numerous authors (e.g., Brogan and Kutner, 1980; Crager, 1987; Laird, 1983; Stanek, 1988; Stein, 1989; Follmann, 1991; Yang and Tsiatis, 2001) have studied these "popular" procedures under various assumptions, yet no general consensus has emerged regarding a preferred approach, providing little guidance for practice.

A further complication facing the data analyst, particularly in lengthy studies, is that of missing follow-up response Y_2 for some subjects. In the ACTG 175 data, for example, although baseline CD4 (Y_1) is available for all 2139 participants, 37% are missing CD4 count at 96±5 weeks (Y_2) due to dropout from the study. A common approach in this situation is to undertake a complete-case analysis, applying one of the above techniques only to the data from subjects for whom both the pre- and posttest responses are observed. In the GEE method, one may in fact include data from all subjects by defining the "multivariate response" for those with missing Y_2 to be simply Y_1, with mean μ_1. However, as is well known, for all of these approaches, unless the data are missing completely at random (Rubin, 1976), which implies that missingness is not associated with any observed or unobserved subject characteristics, these strategies may yield biased inference on β.

Often, baseline demographic and physiologic characteristics X_1, say, are collected on each participant. Moreover, during the intervening period from baseline to follow-up, additional covariate information X_2, say, including intermediate measures of the response, may be obtained. In ACTG 175, at baseline CD4 count (Y_1) and covariates (X_1), including weight; age; indicators of intravenous drug use, HIV symptoms, prior experience with antiretroviral therapy, hemophilia, sexual preference, gender and race; CD8 count (another measure of immune status); and Karnofsky score (an index that reflects a subject's ability to perform

activities of daily living) were recorded for each participant. In addition, CD4 and CD8 counts and treatment status (on/off assigned treatment) were recorded intermittently between baseline and 96 ± 5 weeks (X_2). Missingness at follow-up is often associated with baseline response and baseline and intermediate covariates, and this relationship may be differential by intervention. For example, HIV-infected patients who are worse off at baseline as suggested by low baseline CD4 count may be more likely to drop out, particularly if they receive the less effective treatment. Moreover, HIV-infected patients may base a decision to drop out on post-baseline intermediate measures of immunologic or virologic status (e.g., CD4 counts), which themselves may reflect the effectiveness of their assigned therapy. Here, the assumption that follow-up is missing at random (MAR; Rubin, 1976), associated only with these observable quantities and not the missing response, may be reasonable.

If one is willing to adopt the MAR assumption, methods that take appropriate account of the missingness should be used to ensure valid inference. A standard approach to missing data problems is maximum likelihood, which in the pretest–posttest setting with Y_2 MAR involves full (parametric) specification of the joint distribution of $V = (X_1, Y_1, X_2, Y_2, Z)$. Alternatively, adaptation of popular estimators such as ANCOVA to handle MAR Y_2 on a case-by-case basis may be possible. Maximum likelihood techniques are known to suffer potential sensitivity to deviations from modeling assumptions, and neither approach has been widely applied by practitioners in the pretest–posttest context.

In summary, although missing follow-up response is commonplace in the pretest–posttest setting, there is no widely accepted or used methodology for handling it. In this paper we demonstrate how a unified framework for pretest–posttest analysis under MAR may be developed by exploiting the results in a landmark paper by Robins, Rotnitzky and Zhao (1994).

1.2 Semiparametric Models, Influence Functions, and Robins, Rotnitzky and Zhao

A popular modeling approach that acknowledges concerns over sensitivity to parametric assumptions is to take a semiparametric perspective. A *semiparametric model* may involve both parametric and nonparametric components, where the nonparametric component represents features on which the analyst is unwilling or unable to make parametric assumptions, and interest may focus on a parametric component or on some functional of the nonparametric component. For example, in a regression context one may adopt a parametric model for the conditional expectation of a continuous response given covariates and seek inference on the model parameters but be uncomfortable assuming the full conditional distribution is normal, instead leaving it unspecified. Under the semiparametric model for the pretest–posttest trial we consider, features of the joint distribution of $V = (X_1, Y_1, X_2, Y_2, Z)$ beyond the independence of (X_1, Y_1) and Z induced by randomization are left unspecified and thus constitute the nonparametric component, and interest focuses on the functional β of this distribution defined in (1). When Y_2 is MAR, this semiparametric view not only offers protection against incorrect assumptions on V, but allows us to exploit the theory of Robins, Rotnitzky and Zhao (1994) to deduce estimators for β. These authors derived an asymptotic theory for inference in general semiparametric models with data MAR that may be used to identify a class of consistent estimators for parametric components or such functionals, as we now outline.

Robins, Rotnitzky and Zhao (1994) restricted attention to estimators that are *regular and asymptotically linear*. Regularity is a technical condition that rules out "pathological" estimators with undesirable local properties (Newey, 1990), such as the "superefficient" estimator of Hodges (e.g., Casella and Berger, 2002, page 515). Generically, an estimator $\widehat{\beta}$ for β $(p \times 1)$ in a parametric or semiparametric statistical model for a random vector W based on i.i.d. data W_i, $i = 1, \ldots, n$, is asympotically linear if it satisfies, for a function $\varphi(W)$,

$$n^{1/2}(\widehat{\beta} - \beta_0) = n^{-1/2} \sum_{i=1}^{n} \varphi(W_i) + o_p(1), \tag{2}$$

where β_0 is the true value of β generating the data, $E\{\varphi(W)\} = 0$, $E\{\varphi^T(W)\varphi(W)\} < \infty$ and expectation is with respect to the true distribution of W. The function $\varphi(W)$ is referred to as the *influence function* of $\widehat{\beta}$, as to a first-order $\varphi(W)$ is the influence of a single observation on $\widehat{\beta}$ in the sense given in Casella and Berger (2002, Section 10.6.4). An estimator that is both regular and asymptotically linear (RAL) with influence function $\varphi(W)$ is consistent and asymptotically normal with asymptotic covariance matrix $E\{\varphi(W)\varphi^T(W)\}$. Although not all consistent estimators need be RAL, almost all reasonable estimators are. For RAL estimators, there exists an influence function $\varphi^{\text{eff}}(W)$ such that $E\{\varphi(W)\varphi^T(W)\} - E\{\varphi^{\text{eff}}(W)\varphi^{\text{eff}T}(W)\}$ is nonnegative definite for any influence function $\varphi(W)$; $\varphi^{\text{eff}}(W)$ is referred to as the *efficient influence function*

and the corresponding estimator is called the *efficient estimator*. In fact, for any regular estimator, asymptotically linear or not, with asymptotic covariance matrix Σ, $\Sigma - E\{\varphi^{\text{eff}}(W)\varphi^{\text{eff}T}(W)\}$ is nonnegative definite; thus, the best estimator in the sense of "smallest" asymptotic covariance matrix is RAL, so that restricting attention to RAL estimators is not a limitation. We use the term "influence function" unqualified to mean the influence function of an RAL estimator.

As indicated by (2), there is a relationship between influence functions and consistent and asymptotically normal estimators; thus, by identifying influence functions, one may deduce corresponding estimators. In missing data problems, Robins, Rotnitzky and Zhao (1994) distinguished between full-data and observed-data influence functions. "Full data" refers to the data that would be observed if there were no missingness; in the pretest–posttest setting the full data are V. Accordingly, full-data influence functions correspond to estimators that could be calculated if full data were available and are hence functions of the full data. "Observed data" refers to the data observed when some components of the full data are potentially missing; hence observed-data influence functions correspond to estimators that can be computed from observed data only and are functions of the observed data. For a general semiparametric model, the pioneering contribution of Robins, Rotnitzky and Zhao (1994) was to characterize the class of all observed-data influence functions when data are MAR, including the efficient influence function, and to demonstrate that observed-data influence functions may be expressed in terms of full-data influence functions. Because for many popular semiparametric models the form of full-data influence functions is known or straightforwardly derived, this provides an attractive basis for identifying estimators when data are MAR, including the efficient one.

In summary, the Robins, Rotnitzky and Zhao (1994) theory provides a series of steps for deducing estimators for a semiparametric model of interest when data are MAR: (1) Characterize the class of full-data influence functions, (2) characterize the observed data under MAR and apply the Robins, Rotnitzky and Zhao theory to obtain the class of observed-data influence functions, including the efficient one and (3) identify observed-data estimators with influence functions in this class. In this paper, for the semiparametric pretest–posttest model when Y_2 is MAR, β is a scalar ($p = 1$), and we carry out each of these steps and show how they lead to closed-form estimators for β suitable for routine practical use. Interestingly, despite the ubiquity of the pretest–posttest study and the simplicity of the model when no data are missing, to our knowledge explicit application of this powerful theory to pretest–posttest inference with data MAR with an eye toward developing practical estimators has not been reported.

1.3 Objectives and Summary

The goals of this paper are twofold. The first main objective is to develop accessible practical strategies for inference on β in a semiparametric pretest–posttest model with follow-up data MAR by using the fundamental theory of Robins, Rotnitzky and Zhao (1994) as described above. Although this theory is well known to experts, many researchers have only passing familiarity with its essential elements. Thus, the second main goal of this paper is to use the pretest–posttest problem as a backdrop to provide a detailed and mostly self-contained demonstration of application of the theory of semiparametric models and the powerful, general Robins, Rotnitzky and Zhao results in a concrete, familiar context. This account hopefully will serve as a resource to researchers and practitioners wishing to appreciate the scope and underpinnings of the Robins, Rotnitzky and Zhao theory by systematically tracing the key concepts and steps involved in its application and explicating how it can lead to practical insight and tools.

In Section 2 we summarize the semiparametric pretest–posttest model and outline how the class of full-data influence functions for estimators for β may be derived. In Section 3 we characterize the observed data when Y_2 is MAR, review the essential Robins, Rotnitzky and Zhao (1994) results and apply them to derive the class of observed-data influence functions. Sections 4 and 5 present strategies for constructing estimators based on observed-data influence functions, and we demonstrate the new estimators by application to the ACTG 175 data in Section 6. Results and practical implications are presented in the main narrative; technical supporting material and details of derivations are given in the Appendix.

As in any missing-data context, validity of the assumption of MAR follow-up response is critical and is best justified with availability of rich baseline and intervening information. We assume throughout that the analyst is well equipped to invoke this assumption.

2. MODEL AND FULL-DATA INFLUENCE FUNCTIONS

2.1 Semiparametric Pretest–Posttest Model

First, consider the full data (no missing posttest response). Suppose each subject $i = 1, \ldots, n$ is ran-

domized to treatment with known probability δ, so $Z_i = 0$ or 1 as i is assigned to control or treatment; in ACTG 175, $\delta = 0.75$. Then Y_{1i} and Y_{2i} are i's pretest and posttest responses (baseline and 96±5 week CD4 in ACTG 175), X_{1i} is i's vector of baseline covariates and X_{2i} is the vector of additional covariates collected on i after intervention but prior to follow-up, which may include intermediate measures of response. Assuming subjects' responses evolve independently, the full data on i are $V_i = (X_{1i}, Y_{1i}, X_{2i}, Y_{2i}, Z_i)$, i.i.d. across i with density $p(v) = p(x_1, y_1, x_2, y_2, z)$; we often suppress the subscript i for brevity. From (1) interest focuses on $\beta = E(Y_2|Z = 1) - E(Y_2|Z = 0) = \mu_2^{(1)} - \mu_2^{(0)}$, $\mu_2^{(1)} = \mu_2 + \beta$ and $\mu_2^{(0)} = \mu_2$; throughout, expectation and variance are with respect to the true distribution of V.

From Section 1.2, under a semiparametric perspective the analyst may be unwilling to make specific parametric assumptions on $p(x_1, y_1, x_2, y_2, z)$ such as normality or equality of variances of Y_1 and Y_2. For example, in HIV research it is customary to assume that CD4 counts are normally distributed on some transformed scale and to carry out analyses on this scale; however, as there is no consensus on an appropriate transformation, methods that do not require this assumption are desirable. Thus, in arguments to deduce the form of full- and observed-data influence functions here and in Sections 3 and 4, we do not impose any specific assumptions beyond independence of (X_1, Y_1) and Z induced by randomization and assumptions on the form of the mechanism governing missingness. As our objective is to outline the salient features of the arguments without dwelling on technicalities, we assume needed moments, derivatives and matrix inverses exist without comment.

2.2 Full-Data Influence Functions

As presented in Section 1.2, our first step in applying the Robins, Rotnitzky and Zhao (1994) theory is to characterize the class of all full-data influence functions for RAL estimators for β; these will be functions of V. This may be accomplished by appealing to the theory of semiparametric models (e.g., Newey, 1990; Bickel, Klaassen, Ritov and Wellner, 1993), which provides a formal framework for characterizing influence functions for RAL estimators in such models, including the efficient influence function. The theory takes a geometric perspective, where, generically, influence functions based on data V for RAL estimators for a p-dimensional parameter or functional β in a statistical model for V are viewed as elements of a particular "space" of mean-zero, p-dimensional functions of V for which there is a certain relationship between the distance of any element of the space from the origin and the covariance matrix of the function. From (2), as the covariance matrix of an influence function is equal to the asymptotic covariance matrix of the corresponding estimator, the search for estimators with small covariance matrices, especially the efficient estimator, may thus be focused on functions in this space and guided by geometric distance considerations.

In Appendix A.1 we first sketch an argument that demonstrates that any RAL estimator has a unique influence function, supporting the premise of working with influence functions. We then review familiar results for fully parametric models and show how they may be regarded from this geometric perspective. Finally, we indicate how this perspective is extended to handle semiparametric models. The key results are a representation of the form of all influence functions for RAL estimators in a particular model and a convenient characterization of the efficient influence function that corresponds to the efficient estimator.

In Appendix A.2 we apply these results to show that all full-data influence functions for estimators for β in the semiparametric pretest–posttest model must be of the form

$$\begin{aligned}(3)\qquad &\left\{\frac{Z(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(Z - \delta)}{\delta}h^{(1)}(X_1, Y_1)\right\} \\ &- \left\{\frac{(1 - Z)(Y_2 - \mu_2)}{1 - \delta}\right. \\ &\qquad \left. - \frac{\{(1 - Z) - (1 - \delta)\}}{1 - \delta}h^{(0)}(X_1, Y_1)\right\},\end{aligned}$$

where $h^{(c)}(X_1, Y_1)$, $c = 0, 1$, are arbitrary functions with $\text{var}\{h^{(c)}(X_1, Y_1)\} < \infty$. Technically, the influence function (3) depends on μ_2 and β through their true values. As is conventional, here and in the sequel we write influence functions as functions of parameters, which highlights their practical use as the basis for deriving estimators, shown in Section 5. From (3), influence functions and hence all RAL estimators for β are functions only of (X_1, Y_1, Y_2, Z) and hence do not depend on X_2. This is intuitively reasonable; because X_2 is a post-intervention covariate, we would not expect it to play a role in estimation of β when Y_2 is observed on all subjects. In Section 3, however, we will observe that when Y_2 is MAR for some subjects, such covariates are important not only for validating the MAR assumption, but for increasing efficiency of estimation of β, as discussed in Robins, Rotnitzky and Zhao (1994, page 848).

The results in Appendix A.2 also show that the efficient influence function, that with smallest variance among all influence functions in class (3), is found by taking

$$h^{(c)}(X_1, Y_1) = E(Y_2|X_1, Y_1, Z = c) - \mu_2^{(c)}, \quad c = 0, 1, \quad \mu_2^{(1)} = \mu_2 + \beta, \quad \mu_2^{(0)} = \mu_2. \tag{4}$$

Thus, if full data were available in ACTG 175, the optimal estimator for β would involve the true regression of 96±5 week CD4 on pretest CD4 and other baseline covariates listed in Section 1.1. Leon, Tsiatis and Davidian (2003) identified class (3) when no intervening covariate X_2 is observed and showed that influence functions for the popular estimators discussed in Section 1.1 are members; for example, the two-sample t test estimator

$$\widehat{\beta}_{2s} = n_1^{-1} \sum_{i=1}^{n} Z_i Y_{2i} - n_0^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_{2i}, \quad n_c = \sum_{i=1}^{n} I(Z_i = c), \quad c = 0, 1, \tag{5}$$

has influence function (3) with $h^{(c)} \equiv 0$, $c = 0, 1$ (see Appendix A.2). Thus, popular estimators are RAL and valid under the semiparametric model, and hence contrary to widespread belief, are consistent and asymptotically normal even if Y_1 and Y_2 are not normally distributed. Leon, Tsiatis and Davidian (2003) also showed that none of the popular estimators has the efficient influence function, suggesting that improved estimators are possible, and proposed estimators based on (4) that offer dramatic efficiency gains over popular methods.

In fact, (3) is the difference of the forms of all influence functions for $\mu_2^{(1)}$ and $\mu_2^{(0)}$, respectively, which may themselves be deduced separately by arguments analogous to those in Appendix A.2. In Appendix A.3 we argue that, for the purposes of identifying observed-data estimators for β, it suffices to identify observed-data influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$ separately. We thus focus for simplicity in Section 3 on estimation of $\mu_2^{(1)}$.

3. OBSERVED DATA INFLUENCE FUNCTIONS

3.1 Semiparametric Pretest–Posttest Model with MAR Follow-Up Response

Suppose now that Y_2 is missing for some subjects, with all other variables observed, and define $R = 0$ or 1 as Y_2 is missing or observed. Then the observed data for subject i are $O_i = (X_{0i}, Y_{1i}, X_{1i}, R_i, R_i Y_{2i}, Z_i)$, i.i.d. across i. We represent the assumption that Y_2 is MAR as

$$\begin{aligned} P(R = 1|X_1, Y_1, X_2, Y_2, Z) &= P(R = 1|X_1, Y_1, X_2, Z) \\ &= \pi(X_1, Y_1, X_2, Z) \geq \varepsilon > 0, \end{aligned} \tag{6}$$

reflecting the reasonable view for a pretest–posttest trial that there is a positive probability of observing Y_2 for any subject. Equation (6) formalizes that missingness does not depend on the unobserved Y_2, but may be associated with baseline and intermediate characteristics and be differential by intervention, the latter highlighted by the equivalent representation

$$\pi(X_1, Y_1, X_2, Z) = Z\pi^{(1)}(X_1, Y_1, X_2) + (1 - Z)\pi^{(0)}(X_1, Y_1, X_2) \tag{7}$$

for $\pi^{(c)}(X_1, Y_1, X_2) = \pi(X_1, Y_1, X_2, c) \geq \varepsilon > 0$, $c = 0, 1$. For ACTG 175, (6) and (7) make explicit the belief that subjects may have been more or less likely to drop out (and hence be missing CD4 at 96±5 weeks) depending on their baseline CD4 and other characteristics as well as intermediate measures of CD4 and CD8 and off-treatment status, where this relationship may be different for patients treated with ZDV only versus the other therapies, but that dropout does not depend on unobserved 96±5 week CD4. Relaxation of the assumption that X_1, Y_1, X_2 are observed for all subjects is discussed in Section 7.

3.2 Complete-Case Analysis and Inverse Weighting

As noted in Section 1.1, a naive approach under these conditions is to conduct a complete-case analysis. For example, using the two-sample t test, estimate β by

$$n_{R1}^{-1} \sum_{i=1}^{n} R_i Z_i Y_{2i} - n_{R0}^{-1} \sum_{i=1}^{n} R_i (1 - Z_i) Y_{2i}, \quad n_{Rc} = \sum_{i=1}^{n} R_i I(Z_i = c), \quad c = 0, 1, \tag{8}$$

the difference in sample means based only on data for subjects with Y_2 observed. Under the semiparametric model, as $E(RZY_2) = E\{ZY_2E(R|X_1, Y_1, X_2, Y_2, Z)\} = E\{ZY_2\pi^{(1)}(X_1, Y_1, X_2)\}$ by (6) and (7), and $E\{RI(Z = 1)\} = E(RZ) = E\{Z\pi^{(1)}(X_1, Y_1, X_2)\}$, the first term in (8) converges in probability to $E\{ZY_2 \cdot \pi^{(1)}(X_1, Y_1, X_2)\}/E\{Z\pi^{(1)}(X_1, Y_1, X_2)\}$, which is not

equal to $E(Y_2|Z=1)=\mu_2^{(1)}$ in general. Similarly, the second term is not consistent for $\mu_2^{(0)}$. Thus, (8) is not a consistent estimator for β in general.

A simple remedy is to incorporate inverse weighting of the complete cases (IWCC; e.g., Horvitz and Thompson, 1952). Here, whereas the estimator for $\mu_2^{(1)}$ in (8) solves $\sum_{i=1}^n R_i Z_i (Y_{2i}-\mu_2^{(1)})=0$, weight each contribution by the inverse of the probability of seeing a complete case; that is, solve $\sum_{i=1}^n R_i Z_i (Y_{2i}-\mu_2^{(1)})/\pi_i^{(1)}(X_{1i}, Y_{1i}, X_{2i})=0$, yielding the estimator for $\mu_2^{(1)}$,

$$n_{RZ(1)}^{-1}\sum_{i=1}^n R_i Z_i Y_{2i}/\pi^{(1)}(X_{1i}, Y_{1i}, X_{2i}),$$
$$n_{RZ(1)}=\sum_{i=1}^n R_i Z_i/\pi^{(1)}(X_{1i}, Y_{1i}, X_{2i}), \tag{9}$$

and analogously for $\mu_2^{(0)}$. It is straightforward to show that such inverse weighting yields consistent estimators for $\mu_2^{(c)}$, $c=0,1$; for example, for (9) ($c=1$), using (6) and (7),

$$E\left\{\frac{RZY_2}{\pi^{(1)}(X_1, Y_1, X_2)}\right\}$$
$$=E\left\{ZY_2\frac{E(R|X_1, Y_1, X_2, Y_2, Z)}{\pi^{(1)}(X_1, Y_1, X_2)}\right\}$$
$$=E(ZY_2)=E\{ZE(Y_2|Z)\}=\delta E(Y_2|Z=1),$$

and similarly $E\{RZ/\pi^{(1)}(X_1, Y_1, X_2)\}=\delta$, so that (9) converges in probability to $E(Y_2|Z=1)=\mu_2^{(1)}$. Subtracting $\mu_2^{(c)}$ and multiplying by $n^{1/2}$ for each of $c=0,1$, the associated influence functions are seen to be

$$\frac{RZ(Y_2-\mu_2^{(1)})}{\delta\pi^{(1)}(X_1, Y_1, X_2)} \quad \text{and}$$
$$\frac{R(1-Z)(Y_2-\mu_2^{(0)})}{(1-\delta)\pi^{(0)}(X_1, Y_1, X_2)}, \tag{10}$$

which have the form of the corresponding full-data influence functions in (3) weighted by $1/\pi^{(c)}$, $c=0,1$, for the complete cases only ($R=1$). The IWCC may be applied to any RAL estimator with influence function in class (3), including popular ones. However, although such simple IWCC leads to consistent inference, methods with greater efficiency are possible.

3.3 The Robins, Rotnitzky and Zhao Theory

As noted in Section 1.2, the pioneering advance of Robins, Rotnitzky and Zhao (1994) was to derive, for a general semiparametric model, the class of all observed data influence functions for estimators for a parameter β under complex forms of MAR and to characterize the efficient influence function. The theory reveals, perhaps not unexpectedly, that there is a relationship between full- and observed-data influence functions and that the latter involve inverse weighting.

Denote the subset of the full data V that is always observed for all subjects as O^*; $O^*=(X_1, Y_1, X_2, Z)$ here. Under MAR, the probability that full data are observed depends only on O^*, which we write as $\pi(O^*)$. Assuming $\pi(O^*)$ is known for now, if $\varphi^F(V)$ is any full-data influence function, Robins, Rotnitzky and Zhao showed that, in general, all observed-data influence functions have the form $R\varphi^F(V)/\pi(O^*)-g(O)$, where $g(O)$ is an arbitrary square-integrable function of the observed data that satisfies $E\{g(O)|V\}=0$. For situations like that here, where a particular subset of V (Y_2) is either missing or not for all subjects, this becomes

$$\frac{R\varphi^F(V)}{\pi(O^*)}-\frac{R-\pi(O^*)}{\pi(O^*)}g(O^*), \tag{11}$$

where $g(O^*)$ is an arbitrary square-integrable function of the data always observed. In (11), the first term has the form of an IWCC full-data influence function; the second term, which has mean zero, depending only on data observed for all subjects, "augments" (e.g., Robins, 1999) the first, which leads to increased efficiency provided that g is chosen judiciously.

3.4 Observed-Data Influence Functions for the Pretest–Posttest Problem

In the special case of the pretest–posttest problem, focusing on estimation of the treatment mean $\mu_2^{(1)}=\mu_2+\beta$, with $O^*=(X_1, Y_1, X_2, Z)$, (3) and (11) immediately imply that the class of all observed-data influence functions for estimators for $\mu_2^{(1)}$ when Y_2 is MAR is

$$\frac{R\{Z(Y_2-\mu_2^{(1)})-(Z-\delta)h^{(1)}(X_1, Y_1)\}}{\delta\pi(X_1, Y_1, X_2, Z)}$$
$$-\frac{R-\pi(X_1, Y_1, X_2, Z)}{\pi(X_1, Y_1, X_2, Z)}g^{(1)}(X_1, Y_1, X_2, Z) \tag{12}$$

for arbitrary $h^{(1)}$ and $g^{(1)}$ such that $\text{var}\{h^{(1)}(X_1, Y_1)\}<\infty$ and $\text{var}\{g^{(1)}(X_1, Y_1, X_2, Z)\}<\infty$. Defining $g^{(1)'}(X_1, Y_1, X_2, Z)=(Z-\delta)h^{(1)}(X_1, Y_1)+\delta g^{(1)}(X_1, Y_1, X_2, Z)$, we may write (12) equivalently in a way

that is convenient in subsequent developments as

$$\begin{aligned}(13)\quad &\frac{RZ(Y_2-\mu_2^{(1)})}{\delta\pi(X_1,Y_1,X_2,Z)}-\frac{(Z-\delta)}{\delta}h^{(1)}(X_1,Y_1)\\ &-\frac{R-\pi(X_1,Y_1,X_2,Z)}{\delta\pi(X_1,Y_1,X_2,Z)}g^{(1)'}(X_1,Y_1,X_2,Z);\end{aligned}$$

there is a one-to-one correspondence between (12) and (13).

As in the full-data problem, it is of interest to identify the optimal choices of $h^{(1)}$ and $g^{(1)}$, or, equivalently, $h^{(1)}$ and $g^{(1)'}$, that is, those that yield the efficient observed-data influence function with smallest variance among all influence functions of form (12) or, equivalently, (13). In Appendix A.4 we show that the optimal choices of $h^{(1)}$ and $g^{(1)'}$ in (13) are

$$\begin{aligned}(14)\quad &h^{\mathrm{eff}(1)}(X_1,Y_1)\\ &\quad=E(Y_2|X_1,Y_1,Z=1)-\mu_2^{(1)},\\ &g^{\mathrm{eff}(1)'}(X_1,Y_1,X_2,Z)\\ &\quad=Z\{E(Y_2|X_1,Y_1,X_2,Z)-\mu_2^{(1)}\}\\ &\quad=Z\{E(Y_2|X_1,Y_1,X_2,Z=1)-\mu_2^{(1)}\}.\end{aligned}$$

The forms $g^{\mathrm{eff}(1)'}$ and $h^{\mathrm{eff}(1)}$ show explicitly how augmentation exploits relationships among variables to gain efficiency. In ACTG 175, then, (14) shows that the optimal estimator for β involves knowledge of the true regressions of 96±5 week CD4 on baseline CD4 and other baseline covariates, and on this baseline information plus post-intervention CD4 and CD8 measures and off-treatment status, respectively.

To develop estimators for practical use with good properties, it is sensible to consider influence functions with form close to that of the efficient influence function. Accordingly, from the expression for $g^{\mathrm{eff}(1)'}$ in (14) and the representation of π in (7), we restrict attention in the sequel to the subclass of (13) with elements of the form, for $g^{(1)'}(X_1,Y_1,X_2,Z)=Zq^{(1)}(X_1,Y_1,X_2)$ for arbitrary square-integrable $q^{(1)}(X_1,Y_1,X_2)$,

$$\begin{aligned}(15)\quad &\psi(X_1,Y_1,X_2,R,RY_2,Z)\\ &\quad=\frac{RZ(Y_2-\mu_2^{(1)})}{\delta\pi^{(1)}(X_1,Y_1,X_2)}-\frac{(Z-\delta)}{\delta}h^{(1)}(X_1,Y_1)\\ &\qquad-\frac{\{R-\pi^{(1)}(X_1,Y_1,X_2)\}Z}{\delta\pi^{(1)}(X_1,Y_1,X_2)}\\ &\qquad\cdot q^{(1)}(X_1,Y_1,X_2).\end{aligned}$$

Equation (15) includes the optimal $g^{(1)'}$, but rules out choices that cannot have the efficient form.

3.5 Estimation of the Missingness Mechanism

The foregoing results take π and, hence, $\pi^{(1)}(X_1,Y_1,X_2)$ to be known, which is unlikely unless Y_2 is missing purposefully by design for some subjects in a way that depends on a subject's baseline and intermediate information. In practice, unknown $\pi^{(1)}$ is often addressed by positing a parametric model for $\pi^{(1)}$; intuition suggests that such a model be correctly specified, although we discuss this further in Section 4.2. For now, then, suppose that a parametric model $\pi^{(1)}(X_1,Y_1,X_2;\gamma)$, say, for γ $(s\times 1)$ has been proposed and is correct, where γ_0 is the true value of γ so that evaluation at γ_0 yields the true probability $\pi^{(1)}(X_1,Y_1,X_2)$. For definiteness, we focus henceforth on the logistic regression model

$$\begin{aligned}(16)\quad &\pi^{(1)}(X_1,Y_1,X_2;\gamma)\\ &\quad=\exp\{d^T(X_1,Y_1,X_2)\gamma\}\\ &\qquad\cdot[1+\exp\{d^T(X_1,Y_1,X_2)\gamma\}]^{-1},\end{aligned}$$

where $d(X_1,Y_1,X_2)$ is a vector of functions of its argument, but a development analogous to that below is possible for other choices (e.g., a probit model). In the ACTG 175 analysis in Section 6 we model the probability of observing CD4 at 96±5 weeks by a logistic function, where $d(X_1,Y_1,X_2)$ includes functions of baseline and intermediate characteristics.

Under these conditions, a natural strategy is to derive an estimator for $\mu_2^{(1)}$ from an influence function of the form (15), assuming that $\pi^{(1)}(X_1,Y_1,X_2)$ is known; estimate γ based on the i.i.d. data $(X_{1i},Y_{1i},X_{2i},R_i,Z_i)$, $i=1,\ldots,n$, and substitute the estimated value for γ in the (correct) parametric model $\pi^{(1)}(X_1,Y_1,X_2;\gamma)$; and estimate $\mu_2^{(1)}$ acting as if $\pi^{(1)}$ were known. Robins, Rotnitzky and Zhao (1994) showed that, for any choice of $h^{(1)}$ and $q^{(1)}$ in (15), as long as an efficient procedure [e.g., maximum likelihood (ML)] is used to estimate γ, the resulting influence function for the estimator for β obtained by this strategy has the form

$$\begin{aligned}(17)\quad &\psi(X_1,Y_1,X_2,R,RY_2,Z)\\ &\qquad+d^T(X_1,Y_1,X_2)A_{(1)}^{-1}\big(b_{q(1)}-b_{(1)}\big)\\ &\qquad\cdot\frac{\{R-\pi^{(1)}(X_1,Y_1,X_2)\}Z}{\delta},\end{aligned}$$

where

$$\begin{aligned}b_{(1)}=E\big[&\big(Y_2-\mu_2^{(1)}\big)\{1-\pi^{(1)}(X_1,Y_1,X_2)\}\\ &\cdot d(X_1,Y_1,X_2)|Z=1\big],\end{aligned}$$

$$b_{q(1)} = E\big[q^{(1)}(X_1, Y_1, X_2)\{1 - \pi^{(1)}(X_1, Y_1, X_2)\} \cdot d(X_1, Y_1, X_2)|Z = 1\big],$$

$$A_{(1)} = E\big[\pi^{(1)}(X_1, Y_1, X_2)\{1 - \pi^{(1)}(X_1, Y_1, X_2)\} \cdot d(X_1, Y_1, X_2)d^T(X_1, Y_1, X_2)|Z = 1\big],$$

and $\pi^{(1)}(X_1, Y_1, X_2)$ is the true probability (i.e., the parametric model evaluated at γ_0). In Appendix A.5 we give the basis for this result. Thus, estimators for $\mu_2^{(1)}$ with influence functions in class (17) may be derived by finding estimators with influence functions in class (15) (so for "γ known" in the context of a correct parametric model for $\pi^{(1)}$) and substituting the ML estimator for γ. Thus, although influence functions of the form (17) are useful for understanding the properties of estimators for $\mu_2^{(1)}$ when γ is estimated, one need only work with influence functions of the form (15) to derive estimators.

When $q^{(1)}(X_1, Y_1, X_2)$ has the efficient form $E(Y_2|X_1, Y_1, X_2, Z = 1) - \mu_2^{(1)}$, $b_{(1)} = b_{q(1)}$. Hence, as long as the parametric model for $\pi^{(1)}$ is correct, even if γ is estimated, the last term in (17) is identically equal to zero, but this will not necessarily be true otherwise. This reflects the general result shown by Robins, Rotnitzky and Zhao (1994) that an estimator derived from the efficient influence function will have the same properties whether the parameters in a (correct) model for the missingness mechanism are known or estimated. For general $h^{(1)}$ and $q^{(1)}$ not necessarily equal to the optimal choices, the theory also implies the seemingly counterintuitive result that, even if γ is known, estimating it anyway can lead to a gain in efficiency; that is, for a specific (nonoptimal) choice of $h^{(1)}$ and $q^{(1)}$, the variance of (17) is at least as small as that of (15). In Appendix A.5 we give a justification of this claim.

3.6 Summary

By a development entirely similar to that above for influence functions for estimators for $\mu_2^{(1)}$, we may obtain similar influence functions for estimators for $\mu_2^{(0)}$. Here, influence functions in a subclass that contains the efficient influence function are of the form (15) with Z, $\pi^{(1)}$, δ, $h^{(1)}$ and $q^{(1)}$ replaced by $1 - Z$, $\pi^{(0)}$, $(1 - \delta)$ and analogous functions $h^{(0)}$ and $q^{(0)}$, respectively, with similar modifications in (17). The efficient influence function has, analogous to (14), $h^{\text{eff}(0)} = E(Y_2|X_1, Y_1, Z = 0) - \mu^{(0)}$ and $q^{\text{eff}(0)} = E(Y_2|X_1, Y_1, X_2, Z = 0) - \mu_2^{(0)}$. To deduce estimators for β when Y_2 is MAR, we derive estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$ from these developments and take their difference, which is justified by the argument in Appendix A.3.

It may be shown that if the true missingness mechanism follows a parametric model $\pi(X_1, Y_1, X_2, Z; \gamma)$, inducing models $\pi^{(c)}(X_1, Y_1, X_2; \gamma)$, $c = 0, 1$, correctly specifying this model and estimating γ by ML from the data for subjects with $Z = 0$ and 1 separately leads to estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$ at least as efficient as those found by estimating γ by fitting $\pi(X_1, Y_1, X_2, Z; \gamma)$ to all the data jointly. We recommend this approach in practice.

4. ESTIMATORS FOR β

4.1 Derivation of Estimators from Influence Functions

As a generic principle, based on (2), to identify an estimator from a given influence function, one sets the sum of terms that have the form of the influence function for each subject $i = 1, \ldots, n$ to zero, regarding the influence function as a function of the parameter of interest, and solves for this parameter, possibly substituting estimators for other unknown quantities. In complex models, particularly when $p > 1$, it may be impossible to solve for the parameter explicitly, and this and additional considerations can lead to computational and other challenges. However, for the simple pretest–posttest model, this strategy straightforwardly leads to closed-form estimators for β, as we now demonstrate.

The form of the efficient influence function is a natural starting point from which to derive estimators with good properties. Thus, focusing on $\mu_2^{(1)}$, applying this strategy to (15) with the optimal choices $h^{(1)}(X_1, Y_1) = E(Y_2|X_1, Y_1, Z = 1) - \mu_2^{(1)}$ and $q^{(1)}(X_1, Y_1, X_2) = E(Y_2|X_1, Y_1, X_2, Y = 1) - \mu_2^{(1)}$, simple algebra yields

$$\begin{aligned}\mu_2^{(1)} = {} & (n\delta)^{-1}\sum_{i=1}^{n}\frac{R_i Z_i Y_{2i}}{\pi^{(1)}(X_{1i}, Y_{1i}, X_{2i})} \\ & - (n\delta)^{-1}\sum_{i=1}^{n}(Z_i - \delta) \\ & \cdot E(Y_{2i}|X_{1i}, Y_{1i}, Z_i = 1) \\ & - (n\delta)^{-1}\sum_{i=1}^{n}\frac{\{R_i - \pi^{(1)}(X_{1i}, Y_{1i}, X_{2i})\}Z_i}{\pi^{(1)}(X_{1i}, Y_{1i}, X_{2i})} \\ & \cdot E(Y_{2i}|X_{1i}, Y_{1i}, X_{2i}, Z_i = 1),\end{aligned} \tag{18}$$

and similarly for $\mu_2^{(0)}$. Thus, to estimate β, one would take the difference of (18) and the analogous expression for $\mu_2^{(0)}$. In practice this is complicated by the fact that $\pi(X_1, Y_1, X_2)$ must be modeled and fitted; moreover, it is evident that suitable regression models for $E(Y_2|X_1, Y_1, Z)$ and $E(Y_2|X_1, Y_1, X_2, Z)$ must be identified and fitted. We discuss strategies for resolving these practical challenges in Section 5.

4.2 Double Robustness

So far we have assumed that postulated models for $\pi^{(c)}$, $c = 0, 1$, are correctly specified. If the postulated model is incorrect, substituting this incorrect model into an influence function of the form (15) or (17) when $c = 1$ with arbitrary $h^{(1)}$ and $q^{(1)}$ yields an expression that need not have mean zero; for example, the leading term in $\psi(X_1, Y_1, X_2, R, RY_2, Z)$ in (15) has expectation zero only if $P(R = 1|X_1, Y_1, X_2, Z = 1) = \pi^{(1)}(X_1, Y_1, X_2)$, the true probability, and similarly for $c = 0$. Because a defining characteristic of an influence function is zero mean, estimators derived under such conditions need no longer be consistent. However, there is an exception when the optimal $h^{(1)}$ and $q^{(1)}$ are used as in (18), which we now describe.

In general, the augmentation in (11) induces the interesting property that estimators derived from (11) will be consistent if either (1) the choice $g(O^*)$ does not correspond to the optimal choice but $\pi(O^*)$ is correctly specified or (2) the optimal choice of $g(O^*)$ is used but $\pi(O^*)$ is misspecified. This property is referred to as *double robustness* (e.g., Scharfstein, Rotnitzky and Robins, 1999, Section 3.2.3; van der Laan and Robins, 2003, Section 1.6).

We may demonstrate the double robustness property for estimators for the pretest–posttest model; for definiteness, consider $\mu_2^{(1)}$. Under option 1, with any arbitrary choices for $h^{(1)}$ and $q^{(1)}$, if the model $\pi^{(1)}(X_1, Y_1, X_2; \gamma)$ corresponds to the true mechanism, that (15) has mean zero is immediate. Thus, even if one models $E(Y_2|X_1, Y_1, Z)$ and $E(Y_2|X_1, Y_1, X_2, Z)$ incorrectly in (18), the resulting estimator still has a corresponding legitimate influence function in class (17) (assuming γ is estimated) and hence is consistent. Conversely, under option 2, suppose $E(Y_2|X_1, Y_1, Z)$ and $E(Y_2|X_1, Y_1, X_2, Z)$ are correctly specified in (18), but $\pi^{(1)}(X_1, Y_1, X_2)$ is specified incorrectly by some $\pi^*(X_1, Y_1, X_2)$, say. Substituting π^* for $\pi^{(1)}$ in (18), it is straightforward to show that the right-hand side converges in probability to $\mu_2^{(1)}$ (see Appendix A.6), suggesting that an estimator based on (18) would still be consistent. In fact, the second term in (18) converges in probability to zero even if $E(Y_2|X_1, Y_1, Z = 1)$ is replaced by any arbitrary function of (X_1, Y_1), so that the double robustness property holds if only $E(Y_2|X_1, Y_1, X_2, Z)$ is correct. Of course, if both $\pi^{(1)}$ and $E(Y_2|X_1, Y_1, X_2, Z)$ are specified incorrectly, we cannot expect (18) to yield consistent inference in general.

As we discuss in Section 5, in practice one must develop and fit models for $\pi^{(c)}$, $E(Y_2|X_1, Y_1, Z)$ and $E(Y_2|X_1, Y_1, X_2, Z)$, so the results above are somewhat idealized. However, if the analyst uses his or her best judgment and efforts to develop these models, the chance of coming very close to specifying at least one of them correctly may be high. The theoretical double robustness property suggests that, by using estimators like (18) based on the efficient influence function, the analyst has some protection against inadvertent misspecification. In our experience, even if both types of models are mildly incorrectly specified, valid inferences may be obtained; if one model is grossly incorrect, that with the mild misspecification error tends to dominate, so that reliable inferences are still possible.

5. PRACTICAL IMPLEMENTATION

To obtain estimators for β based on (18) and the analogous expression for $\mu_2^{(0)}$ suitable for practice, $\pi^{(c)}(X_1, Y_1, X_2)$, $E(Y_2|X_1, Y_1, Z = c)$ and $E(Y_2|X_1, Y_1, X_2, Z = c)$, $c = 0, 1$, must be modeled and fitted. Given parametric models $\pi^{(c)}(X_1, Y_1, X_2; \gamma)$, if γ is estimated by ML separately from the data for $Z = 0$ and 1 as at the end of Section 3.6, yielding estimators $\widehat{\gamma}^{(c)}$, $c = 0, 1$, we may form estimated probabilities $\widehat{\pi}_i^{(c)} = \pi^{(c)}(X_{0i}, Y_{1i}, X_{1i}; \widehat{\gamma}^{(c)})$, say. Similarly, given fits of some regression models $E(Y_2|X_1, Y_1, Z = c)$ and $E(Y_2|X_1, Y_1, X_2, Z = c)$, we may obtain predicted values $\widehat{e}_{h(c)i}$ and $\widehat{e}_{q(c)i}$, $c = 0, 1$, say, for $E(Y_{2i}|X_{1i}, Y_{1i}, Z_i = c)$ and $E(Y_{2i}|X_{1i}, Y_{1i}, X_{2i}, Z_i = c)$, respectively. Letting $\widehat{\delta} = n_1/n$, substituting in (18) and its analog for $c = 0$ then yields the estimator $\widehat{\beta} = \widehat{\mu}_2^{(1)} - \widehat{\mu}_2^{(0)}$, where

$$\widehat{\mu}_2^{(1)} = n_1^{-1}\left\{\sum_{i=1}^{n} R_i Z_i Y_{2i}/\widehat{\pi}_i^{(1)} - \sum_{i=1}^{n}(Z_i - \widehat{\delta})\widehat{e}_{h(1)i} - \sum_{i=1}^{n}\big(R_i - \widehat{\pi}_i^{(1)}\big)Z_i\widehat{e}_{q(1)i}/\widehat{\pi}_i^{(1)}\right\}$$

and

$$\widehat{\mu}_2^{(0)} = n_0^{-1}\Bigg\{\sum_{i=1}^{n} R_i(1-Z_i)Y_{2i}/\widehat{\pi}_i^{(0)} + \sum_{i=1}^{n}(Z_i-\widehat{\delta})\widehat{e}_{h(0)i} - \sum_{i=1}^{n}(R_i-\widehat{\pi}_i^{(0)})(1-Z_i)\widehat{e}_{q(1)i}/\widehat{\pi}_i^{(0)}\Bigg\}.$$

Intuitively, replacing the unknown quantities in (18) and its analog for $c=0$ by consistent estimators should not alter the implications for consistency of $\widehat{\beta}$ discussed earlier. We now review considerations involved in obtaining $\widehat{\pi}_i^{(c)}$, $\widehat{e}_{h(c)i}$ and $\widehat{e}_{q(c)i}$, $c=0,1$.

A natural approach to modeling $E(Y_2|X_1, Y_1, Z=c)$ and $E(Y_2|X_1, Y_1, X_2, Z=c)$ is to adopt parametric models based on usual regression considerations. For example, in ACTG 175, Y_2 = CD4 at 96±5 weeks is a continuous measurement, suggesting that standard linear regression models may be used. Because of the assumption of MAR, $E(Y_2|X_1, Y_1, X_2, Z, R)$ does not depend on R; thus, $E(Y_2|X_1, Y_1, X_2, Z) = E(Y_2|X_1, Y_1, X_2, Z, R=1)$, implying that this model may be postulated and fitted based only on the complete cases. Thus, standard techniques for model selection and diagnostics may be applied to the data from subjects with $R=1$. For example, inspection of plots like those in Figure 1, which shows only data for subjects for whom CD4 at 96±5 weeks is observed, may be used. Figure 1 suggests that such reasonable models might include both linear and quadratic terms in Y_1 = baseline CD4.

Considerations for developing and fitting models for $E(Y_2|X_1, Y_1, Z=c)$ are trickier. Ideally, the chosen model for this quantity must be compatible with that for $E(Y_2|X_1, Y_1, X_2, Z)$, as $E(Y_2|X_1, Y_1, Z) = E\{E(Y_2|X_1, Y_1, X_2, Z)|X_1, Y_1, Z\}$. Several practical strategies are possible, although none is guaranteed to achieve this property and hence yield the efficient estimators for $\mu_2^{(c)}$, $c=0,1$. One approach is to adopt a model directly for $E(Y_2|X_1, Y_1, Z)$ that is likely "close enough" to be "approximately compatible." For example, if $E(Y_2|X_1, Y_1, X_2, Z)$ is a linear model in functions of (X_1, Y_1, X_2), one may be comfortable with a linear model for $E(Y_2|X_1, Y_1, Z)$ that includes the same functions of X_1, Y_1. We demonstrate this ad hoc strategy for the ACTG 175 data in Section 6. If all of X_1, Y_1, X_2, Y_2 are continuous, assuming joint normality may be a reasonable approximation, in which case standard results may be used to deduce both models. Alternatively, one might use the relationship $E(Y_2|X_1, Y_1, Z) = E\{E(Y_2|X_1, Y_1, X_2, Z)|X_1, Y_1, Z\}$. For example, for low-dimensional X_2, a distributional model for $X_2|X_1, Y_1, Z$ might be fitted based on the $(X_{1i}, Y_{1i}, X_{2i}, Z_i)$, $i=1,\ldots,n$, which are observed for all subjects; integration with respect to this model would yield the desired conditional quantities for $c=0,1$. For univariate binary X_2,

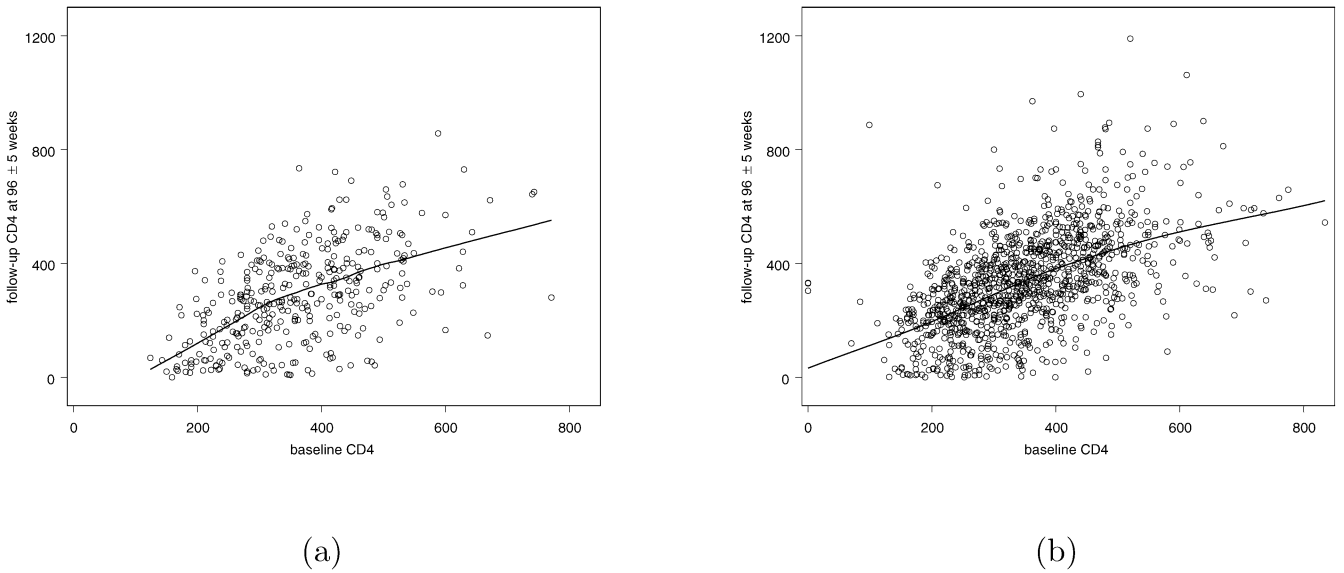

FIG. 1. *CD4 counts after* 96±5 *weeks versus baseline CD4 counts for complete cases for* (a) *ZDV alone and* (b) *the combination of ZDV+ddI, ZDV+ddC or ddI alone, ACTG* 175. *Solid lines were obtained using the Splus function* `loess()` (*Cleveland, Grosse and Shyu*, 1993).

a logistic model for $P(X_2 = 1|X_1, Y_1, Z)$ may be used; this is straightforward, but could be more challenging for mixed continuous and discrete and/or high-dimensional X_2. Instead, one might invoke an empirical approximation, for example, obtaining the predicted value $\widehat{e}_{h(c)i}$, $c = 0, 1$, for each i by averaging estimates of $E(Y_{2i}|X_{1i}, Y_{1i}, X_{2j}, Z_i = c)$ over subjects j that share the same values for (X_1, Y_1, Z) as i, which would likely be feasible only in specialized circumstances. A cruder version would be to average over all X_{2j} for j in the same group as i; this would yield the desired result only if X_2 is conditionally independent of (X_1, Y_1) given Z.

A further complication is that, for any chosen model for $E(Y_2|X_1, Y_1, Z)$, it is no longer appropriate to fit the model based on the complete cases only. Ideally this fitting should be carried out by a procedure that accounts for the fact that Y_2 is MAR, such as an IWCC version of standard regression techniques. However, if the model is an approximation anyway, complete-case-only fitting may not be seriously detrimental. Even if the fit of the chosen model is not consistent for that model, the discussion of double robustness in Section 4.2 suggests that the resulting estimators $\widehat{\mu}_2^{(c)}$ and hence $\widehat{\beta}$ should be consistent regardless.

Another approach would be to use nonparametric smoothing to estimate $E(Y_2|X_1, Y_1, X_1, Z)$ and $E(Y_2|X_1, Y_1, Z)$ and obtain predicted values $\widehat{e}_{q(c)i}$ and $\widehat{e}_{h(c)i}$, for example, locally weighted polynomial smoothing (Cleveland, Grosse and Shyu, 1993) or generalized additive modeling (Hastie and Tibshirani, 1990). Ideally, smoothing for $E(Y_2|X_1, Y_1, Z)$ should be modified to take into account that Y_2 is MAR, although this may not be critical by double robustness. Instead, an estimate of $E(Y_2|X_1, Y_1, Z)$ could be derived from integration of the nonparametric fit of $E(Y_2|X_1, Y_1, X_2, Z)$. Feasibility of smoothing might be limited in high dimensions.

However one approaches developing and fitting models for $E(Y_2|X_1, Y_1, X_2, Z = c)$ and $E(Y_2|X_1, Y_1, Z = c)$, $c = 0, 1$, we have found that it may be advantageous, at least for large n, to fit separate models for $c = 0, 1$. We also recommend including in all four models the same functions of components of X_1, Y_1 and X_2 (if appropriate) if they were found to be important in any one model, as it may be prudent to overmodel rather than undermodel.

Similarly, standard techniques for parametric binary regression may be used to fit models $\pi^{(c)}(X_1, Y_1, X_2; \gamma)$ for each $c = 0, 1$, as all n subjects will have the requisite data. We recommend including in these models all covariates found to be important in any of the regression models above, as it may be shown that including covariates in this model that are correlated with Y_2, even if they are not associated with missingness, can lead to gains in efficiency. Lunceford and Davidian (2004) demonstrated this phenomenon in a simple related setting. Thus, we suggest developing this model after building and fitting of the models for $E(Y_2|X_1, Y_1, X_2, Z = c)$ and $E(Y_2|X_1, Y_1, Z = c)$ are complete.

Theoretically, if all of these models are correctly specified, then $\widehat{\beta}$ should be efficient in the sense described earlier. For parametric regression models for $E(Y_2|X_1, Y_1, X_2, Z = c)$ and $E(Y_2|X_1, Y_1, Z = c)$, although additional regression parameters must be estimated because of the geometry, there is no effect asymptotically; a similar phenomenon for nonparametric estimation of these quantities is suggested by the results of Newey (1990, pages 118–119) as long as this is at a rate faster than $n^{-1/4}$. The double robustness property discussed in Section 4.2 ensures that consistent estimators for β and $\mu_2^{(c)}$, $c = 0, 1$, will be obtained as long as either set of models is correct; however, efficiency is no longer guaranteed.

The asymptotic variance of $\widehat{\beta}$ is obtained from the expectation of the square of the difference of (15) and the analogous control influence function, given by

$$
\begin{aligned}
&E\left\{\frac{(Y_2-\mu_2^{(1)})^2}{\pi^{(1)}(X_1,Y_1,X_2)\delta}\Big|Z=1\right\}\\
&\quad+E\left\{\frac{(Y_2-\mu_2^{(0)})^2}{\pi^{(0)}(X_1,Y_1,X_2)(1-\delta)}\Big|Z=0\right\}\\
&\quad-\delta(1-\delta)\\
&\qquad\cdot E\Bigg[\left\{\frac{E(Y_2|X_1,Y_1,Z=1)-\mu_2^{(1)}}{\delta}\right.\\
&\qquad\qquad\left.+\frac{E(Y_2|X_1,Y_1,Z=0)-\mu_2^{(0)}}{1-\delta}\right\}^2\Bigg]\\
&\quad-\sum_{c=0,1}\left(\frac{I(c=1)}{\delta}+\frac{I(c=0)}{1-\delta}\right)\\
&\qquad\cdot E\Bigg[\frac{1-\pi^{(c)}(X_1,Y_1,X_2)}{\pi^{(c)}(X_1,Y_1,X_2)}\\
&\qquad\qquad\cdot\left\{E(Y_2|X_1,Y_1,X_2,Z=c)\right.\\
&\qquad\qquad\left.-\mu_2^{(c)}\right\}^2\Bigg].
\end{aligned}
\tag{19}
$$

Implicit here is the assumption that the models for $\pi^{(c)}(X_1, Y_1, X_2)$, $E(Y_2|X_1, Y_2, X_2, Z = c)$ and $E(Y_2|$

$X_1, Y_1, Z = c$) are correct. Equation (19) may be estimated by replacing the first two terms by $(\widehat{\delta}\widehat{n}_{RZ(1)})^{-1} \cdot \sum_{i=1}^n R_i Z_i (Y_{2i} - \widehat{\mu}_2^{(1)})^2/\widehat{\pi}_i^{(1)2}$ and $\{(1-\widehat{\delta})\widehat{n}_{RZ(0)}\}^{-1} \cdot \sum_{i=1}^n R_i (1-Z_i)(Y_{2i} - \widehat{\mu}_2^{(0)})^2/\widehat{\pi}_i^{(0)2}$, where $\widehat{n}_{RZ(c)} = \sum_{i=1}^n R_i I(Z_i = c)/\widehat{\pi}_i^{(c)}$, and replacing the remaining terms by sample averages with estimates substituted for needed quantities. Alternatively, $\text{var}(\widehat{\beta})$ may be estimated by $\sum_{i=1}^n \widehat{\varphi}_i^2/n^2$, corresponding to the so-called sandwich technique, where $\widehat{\varphi}_i$ is the difference of the influence functions with estimates substituted, that is,

$$\begin{aligned}\widehat{\varphi}_i = {} & \frac{R_i Z_i (Y_{2i} - \widehat{\mu}_2^{(1)})}{\widehat{\delta}\widehat{\pi}_i^{(1)}} - \frac{(Z_i - \widehat{\delta})(\widehat{e}_{h(1)i} - \widehat{\mu}_2^{(1)})}{\widehat{\delta}} \\ & - \frac{(R_i - \widehat{\pi}_i^{(1)}) Z_i (\widehat{e}_{q(1)i} - \widehat{\mu}_2^{(1)})}{\widehat{\delta}\widehat{\pi}_i^{(1)}} \\ & - \frac{R_i (1-Z_i)(Y_{2i} - \widehat{\mu}_2^{(0)})}{(1-\widehat{\delta})\widehat{\pi}_i^{(0)}} \\ & + \frac{(Z_i - \widehat{\delta})(\widehat{e}_{h(0)i} - \widehat{\mu}_2^{(0)})}{(1-\widehat{\delta})} \\ & - \frac{(R_i - \widehat{\pi}_i^{(0)})(1-Z_i)(\widehat{e}_{q(1)i} - \widehat{\mu}_2^{(0)})}{(1-\widehat{\delta})\widehat{\pi}_i^{(0)}}.\end{aligned}$$

If $E(Y_2|X_1, Y_2, X_2, Z)$ and $E(Y_2|X_1, Y_1, Z)$ are mismodeled, the influence function of $\widehat{\mu}_2^{(1)}$ would instead be of the form (17) to account for estimation of γ, and similarly for $\widehat{\mu}_2^{(0)}$. Although technically then the above formulae would seem to require modification, we have extensive empirical evidence to suggest that they yield reliable estimates of precision if incorrect models are used.

6. TREATMENT EFFECT IN ACTG 175

We now apply the proposed methods to the data from ACTG 175, where β is the difference in mean CD4 count at 96±5 weeks for subjects receiving ZDV (control) and those receiving any of the other three therapies (treatment), so that $\delta = 0.75$. The analysis here is not definitive, but is meant to illustrate the typical steps in an analysis based on these techniques.

Following Section 5, we begin by modeling $E(Y_2|X_1, Y_1, X_2, Z = c)$, $c = 0, 1$. As reviewed in Section 1.1, X_1 contains 11 baseline covariates in addition to baseline CD4 (Y_1). For X_2, we considered three covariates available for all subjects: CD4 at 20±5 weeks postrandomization, CD8 at 20±5 weeks and an indicator of whether the subject went off his/her assigned treatment prior to 96 weeks; reasons could include death, dropout or other patient or physician decisions. Because of the high dimension of X_1 and the fact that both X_1 and X_2 contain a mixture of continuous and discrete variables, we considered parametric linear regression modeling. Based on the 1342 of $n = 2139$ subjects that are complete cases, standard model selection techniques indicate that weight, indicators of HIV symptoms and prior antiretroviral therapy, Karnofsky score, CD8 count and CD4 count (linear and quadratic terms in CD4 and CD8) at baseline, CD8 and CD4 count at 20±5 weeks (linear and quadratic terms), and off-treatment status are associated with Y_2 = CD4 count at 96±5 weeks in one or both treatment groups. Thus, we fit separately for $c = 0, 1$, the models

$$\begin{aligned}& E(Y_2|X_1, Y_1, X_2, Z = c) \\ & \quad = \alpha_0^{(c)} + \alpha_1^{(c)}\text{wt} + \alpha_2^{(c)}\text{HIV} + \alpha_3^{(c)}\text{prior} \\ & \qquad + \alpha_4^{(c)}\text{Karn} + \alpha_5^{(c)}\text{CD8}_0 + \alpha_6^{(c)}\text{CD8}_0^2 \\ & \qquad + \alpha_7^{(c)}\text{CD4}_0 + \alpha_8^{(c)}\text{CD4}_0^2 + \alpha_9^{(c)}\text{CD8}_{20} \\ & \qquad + \alpha_{10}^{(c)}\text{CD8}_{20}^2 + \alpha_{11}^{(c)}\text{CD4}_{20} + \alpha_{12}^{(c)}\text{CD4}_{20}^2 \\ & \qquad + \alpha_{13}^{(c)}\text{offtrt}\end{aligned} \tag{20}$$

by ordinary least squares (OLS), obtaining predicted values $\widehat{e}_{q(c)i}$, $c = 0, 1$ for each $i = 1, \ldots, n$. Adopting the ad hoc strategy in Section 5, we directly modeled $E(Y_2|X_1, Y_1, Z = c)$, $c = 0, 1$, by including the same terms in X_1 and Y_1 as in (20), that is,

$$\begin{aligned}& E(Y_2|X_1, Y_1, Z = c) \\ & \quad = \alpha_0^{(c)} + \alpha_1^{(c)}\text{wt} + \alpha_2^{(c)}\text{HIV} + \alpha_3^{(c)}\text{prior} \\ & \qquad + \alpha_4^{(c)}\text{Karn} + \alpha_5^{(c)}\text{CD8}_0 + \alpha_6^{(c)}\text{CD8}_0^2 \\ & \qquad + \alpha_7^{(c)}\text{CD4}_0 + \alpha_8^{(c)}\text{CD4}_0^2,\end{aligned}$$

again fitting the model for each c by OLS and obtaining predicted values $\widehat{e}_{h(c)i}$, $c = 0, 1$ for all n subjects. Finally, based on standard techniques for logistic regression and the guidelines in Section 5, we arrived at

$$\begin{aligned}& \text{logit}\,\pi^{(c)}(X_1, Y_1, X_2; \gamma^{(c)}) \\ & \quad = \gamma_0^{(c)} + \gamma_1^{(c)}\text{wt} + \gamma_2^{(c)}\text{HIV} + \gamma_3^{(c)}\text{prior} \\ & \qquad + \gamma_4^{(c)}\text{Karn} + \gamma_5^{(c)}\text{CD8}_0 + \gamma_6^{(c)}\text{CD8}_0^2 \\ & \qquad + \gamma_7^{(c)}\text{CD4}_0 + \gamma_8^{(c)}\text{CD4}_0^2 + \gamma_9^{(c)}\text{CD8}_{20} \\ & \qquad + \gamma_{10}^{(c)}\text{CD8}_{20}^2 + \gamma_{11}^{(c)}\text{CD4}_{20} + \gamma_{12}^{(c)}\text{CD4}_{20}^2 \\ & \qquad + \gamma_{13}^{(c)}\text{offtrt}, \quad c = 0, 1,\end{aligned}$$

TABLE 1
Treatment effect estimates for 96±5 *week CD*4 *counts for ACTG* 175

	Estimate	SE
New method	57.24	10.20
IWCC	54.69	11.79
ANCOVA	64.54	9.33
Paired t test	67.14	9.23

NOTE. Standard errors for the new method and the IWCC estimate were obtained using the sandwich approach. ANOVA denotes ordinary analysis of covariance with no interaction term. Standard errors for the popular estimators based on complete cases were obtained from standard formulæ.

where γ was estimated separately by ML for each group.

Table 1 shows the estimate of $\widehat{\beta}$ and the estimated standard error obtained by the sandwich technique, and appears to provide strong evidence that mean CD4 at 96±5 weeks is higher in the treatment group relative to the control. Table 1 also presents the estimate of β obtained via the IWCC method. The IWCC estimated standard error is larger than that for the proposed methods, consistent with the implication of the theory that incorporation of baseline and intervening covariate information should improve precision. For comparison, Table 1 shows estimates of β obtained by the two most popular approaches in practice based on the complete cases only. These results suggest there may be nonnegligible bias associated with these naive methods, in this case suggesting an overly optimistic treatment difference.

7. DISCUSSION

We have shown how the theory developed by Robins, Rotnitzky and Zhao (1994) may be applied to the ubiquitous pretest–posttest problem to deduce analysis procedures that take appropriate account of MAR follow-up data, yield consistent inferences and lead to efficiency gains over simpler methods by exploiting auxiliary covariate information. This perspective provides a general framework for pretest–posttest analysis with missing data that illuminates how relationships among variables play a role in both accounting for missingness and enhancing precision, thus offering the analyst guidance for selecting appropriate methods in practice. We hope that this explicit, detailed demonstration of this theory in a familiar context will help researchers who are not well versed in its underpinnings appreciate the fundamental concepts and how the theoretical results may be translated into practical methods.

We have carried out extensive simulations that show that the proposed methods lead to consistent inference and considerable efficiency gains over simpler methods such as IWCC estimators; a detailed account is available at http://www4.stat.ncsu.edu/~davidian.

We considered the situation where only follow-up response is potentially missing; baseline and intermediate covariates are assumed observable for all subjects. In some settings covariate information in the period between baseline and follow-up may be censored due to dropout, leading to only partially observed X_2. The development may be extended to this case via the Robins, Rotnitzky and Zhao (1994) theory and is related to that for causal inference for time-dependent treatments, requiring assumptions similar to those of sequential randomization identified by Robins (e.g., Robins, 1999; van der Laan and Robins, 2003).

Although our presentation is in the context of the pretest–posttest study, it is evident that the results are equally applicable to the problem of comparing two means in a randomized study with adjustment for baseline covariates to improve efficiency, as discussed, for example, by Koch et al. (1998), because the pretest response Y_1 may be viewed as simply another baseline covariate. Thus, the developments also clarify how such optimal adjustment should be carried out to achieve efficient inferences on a difference in means in this setting; moreover, they provide a systematic approach to accounting for missing response.

APPENDIX

A.1. INFLUENCE FUNCTIONS AND SEMIPARAMETRIC THEORY

Correspondence between influence functions and RAL estimators. Before we describe semiparametric theory, we sketch an argument that more fully justifies why working with influence functions is informative for identifying (RAL) estimators. It is straightforward to show by contradiction that an asymptotically linear estimator [i.e., an estimator satisfying (2)] has a unique (almost surely) influence function. In the notation in (2), if this were not the case, there would exist another influence function $\varphi^*(W)$ with $E\{\varphi^*(W)\} = 0$ that also satisfies (2). If (2) holds for both $\varphi(W)$ and $\varphi^*(W)$, it must be that $A = n^{-1/2}\sum_{i=1}^n\{\varphi(W_i) - \varphi^*(W_i)\} = o_p(1)$. Whereas the W_i are i.i.d., A converges in distribution to a normal random vector with mean zero and covariance matrix $\Sigma = E[\{\varphi(W) - \varphi^*(W)\}\{\varphi(W) - \varphi^*(W)\}^T]$. Whereas this limiting distribution is $o_p(1)$, it must be that $\Sigma = 0$, implying $\varphi(W) = \varphi^*(W)$ almost surely.

Parametric model. We begin by considering fully parametric models. Formally, a parametric model for data V is characterized by all densities $p(v)$ in a class $\mathcal{P}$ indexed by a q-dimensional parameter θ, so that $p(v) = p(v, \theta) \in \mathcal{P}$ is fully specified by θ. Suppose that interest focuses on a parameter β in this model. In the most familiar case, θ may be partitioned explicitly as $\theta = (\beta^T, \eta^T)^T$ for β $(p \times 1)$ and η $(r \times 1)$, $r = q - p$, so that η is a nuisance parameter, and we may write $p(v, \beta, \eta)$. Alternatively, $\beta = \beta(\theta)$ may be some function of θ, and identifying the nuisance parameter may be less straightforward, but the principles are the same. For simplicity, whereas β in the pretest–posttest problem is a scalar, we restrict attention to $p = 1$.

Maximum likelihood estimator in a parametric model. For definiteness, consider the case where $\theta = (\beta, \eta^T)^T$. Define as usual the score vector $S_\theta(V, \theta) = \{S_\beta(V, \theta), S_\eta^T(V, \theta)\}^T = [\partial/\partial\beta \{\log p(V, \beta, \eta)\}, \partial/\partial\eta^T \{\log p(V, \beta, \eta)\}]^T$ and let $\theta_0 = (\beta_0, \eta_0^T)^T$ be the true value of θ. Then $E\{S_\theta(V, \theta_0)\} = 0$ and the information matrix is

$$\mathcal{I}(\theta_0) = E\{S_\theta(V, \theta_0) S_\theta^T(V, \theta_0)\} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\eta} \\ \mathcal{I}_{\beta\eta}^T & \mathcal{I}_{\eta\eta} \end{pmatrix}, \quad \mathcal{I}_{\eta\eta}\ (r \times r),\ \mathcal{I}_{\beta\eta}\ (1 \times r),$$

where expectation is with respect to the true density $p(v, \beta_0, \eta_0)$. Writing $\widehat{\theta} = (\widehat{\beta}, \widehat{\eta}^T)^T$ to denote the maximum likelihood estimator for θ found by maximizing $\sum_{i=1}^n \log p(v_i, \beta, \eta)$, it is well known (e.g., Bickel et al., 1993, Section 2.4) that, under regularity conditions,

$$\text{(A.1)} \qquad n^{1/2}(\widehat{\beta} - \beta_0) = n^{-1/2} \sum_{i=1}^n \varphi^{\text{eff}}(V_i) + o_p(1),$$

$$\text{(A.2)} \qquad \varphi^{\text{eff}}(V) = \mathcal{I}_{\beta\beta\bullet\eta}^{-1} \{S_\beta(V, \theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1} S_\eta(V, \theta_0)\},$$

$$\mathcal{I}_{\beta\beta\bullet\eta} = \mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}\mathcal{I}_{\beta\eta}^T,$$

so that $E\{\varphi^{\text{eff}}(V)\} = 0$ and $\widehat{\beta}$ is RAL with influence function $\varphi^{\text{eff}}(V)$. Whereas $S^{\text{eff}}(V) = S_\beta(V, \theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1} S_\eta(V, \theta_0)$ has variance $\mathcal{I}_{\beta\beta\bullet\eta}$, $\widehat{\beta}$ is consistent and asymptotically normal with asymptotic variance $E\{\varphi^{\text{eff}}(V)\varphi^{\text{eff}T}(V)\} = 1/\mathcal{I}_{\beta\beta\bullet\eta}$, the well-known Cramér–Rao lower bound, the smallest possible variance for (regular) estimators for β. Thus, $\widehat{\beta}$ is the efficient estimator and, accordingly, $S^{\text{eff}}(V)$ is often called the *efficient score*. Evidently $\varphi^{\text{eff}}(V)$ is the efficient influence function, and these familiar results emphasize the connection between efficiency and the score vector.

We are now in position to place these results in a geometric context. Our discussion of this geometric construction for both parametric and semiparametric models is not meant to be rigorous and complete, but serves only to highlight the crucial elements.

Hilbert space. A *Hilbert space* $\mathcal{H}$ is a linear vector space, so that $ah_1 + bh_2 \in \mathcal{H}$ for $h_1, h_2 \in \mathcal{H}$ and any real a, b, equipped with an inner product; see Luenberger (1969, Chapter 3). The key feature that underlies the geometric perspective is that influence functions based on data V for estimators for a p-dimensional parameter β in a statistical model may be viewed as elements in the particular Hilbert space $\mathcal{H}$ of all p-dimensional, mean-zero random functions $h(V)$ such that $E\{h^T(V)h(V)\} < \infty$, with inner product $E\{h_1^T(V)h_2(V)\}$ for $h_1, h_2 \in \mathcal{H}$ and corresponding norm $\|h\| = [E\{h^T(V)h(V)\}]^{1/2}$, measuring distance from $h \equiv 0$. Thus, the geometry of Hilbert spaces provides a unified framework for deducing results with regard to influence functions in both parametric and semiparametric models.

Some general results concerning Hilbert spaces are important. For any linear subspace M of $\mathcal{H}$, the set of all elements of $\mathcal{H}$ *orthogonal* to those in M, denoted $M^\perp$ (i.e., such that if $h_1 \in M$ and $h_2 \in M^\perp$, the inner product of h_1, h_2 is zero), is also a linear subspace of $\mathcal{H}$. Moreover, for two linear subspaces M and N, $M \oplus N$ is the *direct sum* of M and N if every element in $M \oplus N$ has a unique representation of the form $m + n$ for $m \in M$, $n \in N$. Intuitively, it is the case that the entire Hilbert space $\mathcal{H} = M \oplus M^\perp$. As we will see momentarily, a further essential concept is the notion of a *projection*. The projection of $h \in \mathcal{H}$ onto a closed linear subspace M of $\mathcal{H}$ is the element in M, denoted by $\Pi(h|M)$, such that $\|h - \Pi(h|M)\| < \|h - m\|$ for all $m \in M$ and the *residual* $h - \Pi(h|M)$ is orthogonal to all $m \in M$; such a projection is unique (e.g., Luenberger, 1969, Section 3.3).

In light of the pretest–posttest problem, we again take $p = 1$. Let θ_0 be the true value of θ.

Geometric perspective on the parametric model. Consider first the case where θ may be partitioned as $\theta = (\beta, \eta^T)^T$, η $(r \times 1)$. Let Λ be the linear subspace of $\mathcal{H}$ that consists of all linear combinations of $S_\eta(V, \theta_0)$ of the form $BS_\eta(V, \theta_0)$, that is, $\Lambda = \{BS_\eta(V, \theta_0)$ for all $(1 \times r)$ $B\}$, the linear subspace of $\mathcal{H}$ spanned by $S_\eta(V, \theta_0)$. Whereas Λ depends

on the score for nuisance parameters, it is referred to as the *nuisance tangent space*. A fundamental result in this case is that all influence functions for RAL estimators for β may be shown to lie in the subspace $\Lambda^\perp$ orthogonal to Λ. Although a proof of this is beyond our scope, it is straightforward to provide an example by demonstrating that the efficient influence function in (A.2) lies in $\Lambda^\perp$. In particular, we must show that $E\{\varphi^{\text{eff}T}(V)BS_\eta(V,\theta_0)\} = E[\{S_\beta(V,\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}S_\eta(V,\theta_0)\}^T BS_\eta(V,\theta_0)]/\mathcal{I}_{\beta\beta\bullet\eta} = 0$ for all B $(1\times r)$. By taking B successively to be a $(1\times r)$ vector with a 1 in one component and 0s elsewhere, this may be seen to be equivalent to showing that $E[\{S_\beta(V,\theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1}S_\eta(V,\theta_0)\}S_\eta^T(V,\theta_0)] = 0$, which follows immediately. Thus, one approach to identifying influence functions for a particular model with $\theta = (\beta,\eta^T)^T$ is to characterize the form of elements in $\Lambda^\perp$ directly.

Alternatively, other representations are possible. For general $p(v,\theta)$, the *tangent space* Γ is the linear subspace of $\mathcal{H}$ spanned by the entire score vector $S_\theta(V,\theta_0)$, where $S_\theta(V,\theta) = \partial/\partial\theta\{\log p(V,\theta)\}$, that is, $\Gamma = \{BS_\theta(V,\theta_0)$ for all $(1\times q)$ $B\}$. We have the following key result.

RESULT A.1. *Representation of influence functions*. All influence functions for (RAL) estimators for β may be represented as $\varphi(V) = \varphi^*(V) + \psi(V)$, where $\varphi^*(V)$ is any influence function and $\psi(V)\in\Gamma^\perp$, the subspace of $\mathcal{H}$ orthogonal to Γ.

This may be shown for general $\beta(\theta)$; we demonstrate when $\theta = (\beta,\eta^T)^T$. In this case, a defining property of influence functions $\varphi(V)$ which is related to regularity is that (1) $E\{\varphi(V)S_\beta(V,\theta_0)\} = 1$ and (2) $E\{\varphi(V)S_\eta^T(V,\theta_0)\} = 0$ $(1\times r)$; the proof is outside our scope here. Given this, we now show that all influence functions can be represented as in Result A.1. First, we demonstrate that if $\varphi(V)$ can be written as $\varphi^*(V)+\psi(V)$, where $\varphi^*(V)$ and $\psi(V)$ satisfy the conditions of Result A.1, then $\varphi(V)$ is an influence function. Letting $\Gamma_\beta = \{BS_\beta(V,\theta_0)$ for all real $B\}$ be the space spanned by the score for β, it may be shown that $\Gamma = \Lambda\oplus\Gamma_\beta$. Thus, if $\psi\in\Gamma^\perp$, $\psi(V)$ is orthogonal to functions in both Λ and Γ_β, so that $E\{\psi(V)S_\beta(V,\theta_0)\} = 0$ and $E\{\psi(V)S_\eta^T(V,\theta_0)\} = 0$ $(1\times r)$. Moreover, because $\varphi^*(V)$ is an influence function, it satisfies properties 1 and 2, whence it follows that $\varphi(V)$ also satisfies properties 1 and 2 and, hence, is itself an influence function. Conversely, we show that if $\varphi(V)$ is an influence function, it can be represented as in Result A.1. If $\varphi(V)$ is an influence function, it must satisfy properties 1 and 2, and, writing $\varphi(V) = \varphi^*(V) + \{\varphi(V) - \varphi^*(V)\}$ for some other influence function $\varphi^*(V)$, it is straightforward to use properties 1 and 2 to show that $\psi(V) = \{\varphi(V) - \varphi^*(V)\}\in\Gamma^\perp$. Thus, in general, by identifying any influence function and $\Gamma^\perp$, one may exploit Result A.1 to characterize all influence functions.

Depending on the particular model and nature of β, one method for characterizing influence functions may be more straightforward than another. When using Result A.1 in models where $\theta = (\beta,\eta^T)^T$, Γ may be most easily determined by finding Λ and Γ_β separately; for general $\beta(\theta)$, Γ may often be identified directly.

From Result A.1, we may also deduce a useful characterization of the efficient influence function $\varphi^{\text{eff}}(V)$ that satisfies $E\{\varphi^2(V)\} - E\{\varphi^{\text{eff}\,2}(V)\}\geq 0$ for all influence functions $\varphi(V)$. Whereas for arbitrary $\varphi(V)$, $\varphi^{\text{eff}}(V) = \varphi(V) - \psi(V)$ for $\psi\in\Gamma^\perp$ and $E\{\varphi^{\text{eff}\,2}(V)\} = \|\varphi-\psi\|$ must be as small as possible, it must be that $\psi = \Pi(\varphi|\Gamma^\perp)$. Thus, we have the following result.

RESULT A.2. *Representation of the efficient influence function*. The function $\varphi^{\text{eff}}(V)$ may be represented as $\varphi(V) - \Pi(\varphi|\Gamma^\perp)(V)$ for any influence function $\varphi(V)$.

In the case $\theta = (\beta,\eta^T)^T$, it is in fact possible to identify explicitly the form of the efficient influence function. Here, the *efficient score* is defined as the residual of the score vector for β after projecting it onto the nuisance tangent space, $S^{\text{eff}}(V,\theta_0) = S_\beta(V,\theta_0) - \Pi(S_\beta|\Lambda)$, and the *efficient influence function* is an appropriately scaled version of S^{eff} given by $\varphi^{\text{eff}}(V) = [E\{S^{\text{eff}\,2}(V,\theta_0)\}]^{-1}S^{\text{eff}}(V,\theta_0)$. It is straightforward to observe that $\varphi^{\text{eff}}(V)$ is an influence function by showing it satisfies properties 1 and 2 above. Specifically, by construction $S^{\text{eff}}\in\Lambda^\perp$, so property 2 holds. This implies $E\{\varphi^{\text{eff}}(V)\Pi(S_\beta|\Lambda)(V)\} = 0$, so that

$$\begin{aligned}&E\{\varphi^{\text{eff}}(V)S_\beta(V,\theta_0)\}\\ &\quad= E\{\varphi^{\text{eff}}(V)S^{\text{eff}}(V,\theta_0)\} + E\{\varphi^{\text{eff}}(V)\Pi(S_\beta|\Lambda)(V)\}\\ &\quad= E\{S^{\text{eff2}}(V,\theta_0)\}E\{S^{\text{eff2}}(V,\theta_0)\} = 1,\end{aligned}$$

demonstrating property 1. That $\varphi^{\text{eff}}(V)$ has the smallest variance among influence functions may be seen by using the fact that all influence functions may be written as $\varphi(V) = \varphi^{\text{eff}}(V) + \psi(V)$ for some $\psi(V)\in\Gamma^\perp$. Because $S_\beta\in\Gamma_\beta$ and $\Pi(S_\beta|\Lambda)\in\Lambda$ are both in Γ, it follows that $E\{\psi(V)\varphi^{\text{eff}}(V)\} = 0$. Thus, $E\{\varphi^2(V)\} = E[\{\varphi^{\text{eff}}(V)+\psi(V)\}^2] = E\{\varphi^{\text{eff}\,2}(V)\} + E\{\psi^2(V)\}$, so that any other influence function $\varphi(V)$ has variance at

least as large as that of $\varphi^{\text{eff}}(V)$, and this smallest variance is immediately seen to be $1/S^{\text{eff}\,2}(V, \theta_0)$.

Finally, we may relate this development to the familiar maximum likelihood results when $\theta = (\beta, \eta^T)^T$. By definition, $\Pi(S_\beta|\Lambda) \in \Lambda$ is the unique element $B_0 S_\eta \in \Lambda$ such that $E[\{S_\beta(V, \theta_0) - B_0 S_\eta(V, \theta_0)\} \cdot B S_\eta(V, \theta_0)] = 0$ for all B $(1 \times r)$. As above, this is equivalent to requiring $E[\{S_\beta(V, \theta_0) - B_0 S_\eta(V, \theta_0)\} \cdot S_\eta^T(V, \theta_0)] = 0$ $(1 \times r)$, implying $B_0 = \mathcal{I}_{\beta\beta}\mathcal{I}_{\eta\eta}^{-1}$. Thus, $\Pi(S_\beta|\Lambda) = \mathcal{I}_{\beta\beta}\mathcal{I}_{\eta\eta}^{-1} S_\eta(V, \theta_0)$ and $S^{\text{eff}}(V, \theta_0) = S_\beta(V_i, \theta_0) - \mathcal{I}_{\beta\eta}\mathcal{I}_{\eta\eta}^{-1} S_\eta(V_i, \theta_0)$, as expected.

For a parametric model, it is usually unnecessary to appeal to the foregoing geometric construction to identify the efficient estimator and influence functions. In contrast, in the more complex case of a semiparametric model such results often may not be derived readily. However, as we now discuss, the geometric perspective may be generalized to semiparametric models, providing a systematic framework for identifying influence functions.

Geometric perspective on the semiparametric model. In its most general form, a semiparametric model for data V is characterized by the class $\mathcal{P}$ of all densities $p\{v, \theta(\cdot)\}$ that depend on an infinite-dimensional parameter $\theta(\cdot)$. Often, analogous to the familiar parametric case, $\theta(\cdot) = \{\beta, \eta(\cdot)\}$, where β is $(p \times 1)$ and $\eta(\cdot)$ is an infinite-dimensional nuisance parameter, and interest focuses on β. For example, in the regression situation in Section 1.2, β specifies a parametric model for the conditional expectation of a response given covariates, and $\eta(\cdot)$ represents all remaining aspects, such as other features of the conditional distribution, that are left unspecified. Alternatively, interest may focus on a functional $\beta\{\theta(\cdot)\}$ of $\theta(\cdot)$. This is the case in the semiparametric pretest–posttest model, where $\theta(\cdot)$ represents all aspects of the distribution of $V = (X_1, Y_1, X_2, Y_2, Z)$ that are left unspecified and β is given in (1).

The key to generalization of the results for parametric models to this setting is the notion of a *parametric submodel*. A parametric submodel is a parametric model contained in the semiparametric model that contains the truth. In the most general case, with densities $p\{v, \theta(\cdot)\}$ and functional of interest $\beta\{\theta(\cdot)\}$, there is a true $\theta_0(\cdot)$ such that $p_0(v) = p\{v, \theta_0(\cdot)\} \in \mathcal{P}$ is the density that generates the data. A parametric submodel is the class of all densities $\mathcal{P}_\xi$ characterized by a finite-dimensional parameter ξ such that $\mathcal{P}_\xi \subset \mathcal{P}$ and the true density $p_0(v) = p\{v, \theta_0(\cdot)\} = p(v, \xi_0) \in \mathcal{P}_\xi$, where the dimension r of ξ varies according to the particular choice of submodel. That is, there exists a density identified by the parameter ξ_0 within the parameter space of the parametric submodel such that $p_0(v) = p(v, \xi_0)$. In Appendix A.2 below, we give an explicit example of parametric submodels in the pretest–posttest setting.

The importance of this concept is that an estimator is an (RAL) estimator for β under the semiparametric model if it is an estimator under every parametric submodel. Thus, the class of estimators for β for the semiparametric model must be contained in the class of estimators for a parametric submodel and, hence, any influence function for the semiparametric model must be an influence function for a parametric submodel. Now, if Γ_ξ is the tangent space for a given submodel $p(v, \xi)$ with score vector $S_\xi(v, \xi) = \partial/\partial\xi \{\log p(v, \xi)\}$, by Result A.1 the corresponding influence functions for estimators for β must be representable as $\varphi(V) = \varphi^*(V) + \gamma(V)$, where $\varphi^*(V)$ is any influence function in the parametric submodel and $\gamma(V) \in \Gamma_\xi^\perp$. Thus, intuitively, defining Γ to be the mean square closure of all parametric submodel tangent spaces [i.e., $\Gamma = \{h \in \mathcal{H}$ such that there exists a sequence of parametric submodels $\mathcal{P}_{\xi_j}$ with $\|h(V) - B_j S_{\xi_j}(V, \xi_{0j})\|^2 \to 0$ as $j \to \infty\}$, where B_j are $(1 \times r_j)$ constant matrices], then it may be shown that Result A.1 holds for semiparametric model influence functions. That is, all influence functions $\varphi(V)$ for estimators for β in the semiparametric model may be represented as $\varphi^*(V) + \psi(V)$, where $\varphi^*(V)$ is any semiparametric model influence function and $\psi(V) \in \Gamma^\perp$. Moreover, Result A.2 also holds: as in the parametric case, the efficient estimator with smallest variance has influence function $\varphi^{\text{eff}}(V)$ and may be represented as $\varphi^{\text{eff}}(V) = \varphi(V) - \Pi(\varphi|\Gamma^\perp)(V)$ for any semiparametric model influence function $\varphi(V)$. In Appendix A.2 we use these results to deduce full-data influence functions for the semiparametric pretest–posttest model.

Although the pretest–posttest model may be handled using the above development, it is worth noting that a framework analogous to the parametric case ensues when $\theta(\cdot) = \{\beta, \eta(\cdot)\}$, so that $p(v) = p\{v, \beta, \eta(\cdot)\}$, with true values $\beta_0, \eta_0(\cdot)$ such that the true density is $p_0(v) = p\{v, \beta_0, \eta_0(\cdot)\} \in \mathcal{P}$. Here, a parametric submodel $\mathcal{P}_{\beta,\xi}$ is the class of all densities characterized by β and finite-dimensional ξ such that $\mathcal{P}_{\beta,\xi} \subset \mathcal{P}$, $p\{v, \beta_0, \eta_0(\cdot)\} = p(v, \beta_0, \xi_0) \in \mathcal{P}_{\beta,\xi}$. As a parametric model, a submodel has a corresponding nuisance tangent space and, as above, because

the class of estimators for β for the semiparametric model must be contained in the class of estimators for a parametric submodel, influence functions for estimators for β for the semiparametric model must lie in a space orthogonal to all submodel nuisance tangent spaces. Thus, defining the semiparametric model nuisance tangent space Λ as the mean square closure of all parametric submodel nuisance tangent spaces, it may be shown that all influence functions for the semiparametric model lie in $\Lambda^{\perp}$. Moreover, the semiparametric model tangent space $\Gamma = \Lambda \oplus \Gamma_{\beta}$, where Γ_{β} is the space spanned by $S_{\beta}\{V, \beta_0, \eta_0(\cdot)\} = \partial/\partial\beta\,[\log p\{V, \beta, \eta_0(\cdot,)\}]$ evaluated at β_0. The semiparametric model efficient score S^{eff} is $S_{\beta}\{V, \beta_0, \eta_0(\cdot)\} - \Pi(S_{\beta}|\Lambda)(V)$ with efficient influence function $\varphi^{\text{eff}}\{V, \beta_0, \eta_0(\cdot)\} = \{E([S^{\text{eff}}\{V, \beta_0, \eta_0(\cdot)\}]^2)\}^{-1} \cdot S^{\text{eff}}\{V, \beta_0, \eta_0(\cdot)\}$. The variance of φ^{eff}, $\{E([S^{\text{eff}}\{V, \beta_0, \eta_0(\cdot)\}]^2)\}^{-1}$, achieves the so-called *semiparametric efficiency bound*, that is, the supremum over all parametric submodels of the Cramér–Rao lower bounds for β.

A.2. DERIVATION OF FULL-DATA INFLUENCE FUNCTIONS

We apply the theory in Appendix A.1 to identify the class of all influence functions $\varphi(V)$ for estimators for β depending on the full data $V = (X_1, Y_1, X_2, Y_2, Z)$ under the semiparametric pretest–posttest model with no assumptions on $p(v)$ beyond independence of (X_1, Y_1) and Z. By Result A.1, these may be written as $\varphi(V) = \varphi^*(V) + \psi(V)$, where $\psi(V) \in \Gamma^{\perp}$ and φ^* is any influence function, so we proceed by identifying a φ^* and characterizing $\Gamma^{\perp}$.

To identify a φ^* under the semiparametric model, consider the two-sample t test estimator $\widehat{\beta}_{2s}$ in (5). Using $n_c/n \xrightarrow{p} \delta^c(1-\delta)^{1-c}$, $c = 0, 1$, $E(ZY_2) = E\{ZE(Y_2|Z)\} = \delta E(Y_2|Z = 1)$ and similarly for $E\{(1-Z)Y_2\}$, $\widehat{\beta}_{2s}$ is clearly consistent under the minimal assumptions on $p(v)$, and from the ensuing expression for $n^{1/2}(\widehat{\beta}_{2s} - \beta)$, writing $\beta = \mu_2^{(1)} - \mu_2^{(0)}$ and using $n_c/n \xrightarrow{p} \delta^c(1-\delta)^{1-c}$, it is straightforward to derive the corresponding influence function

$$\varphi^*(V) = Z\big(Y_2 - \mu_2^{(1)}\big)/\delta - (1-Z)\big(Y_2 - \mu_2^{(0)}\big)/(1-\delta), \tag{A.3}$$

where we write this as a function of $\mu_2^{(0)}$ and $\mu_2^{(1)}$ following the convention noted after (3).

To find $\Gamma^{\perp}$, we consider the class $\mathcal{P}$ of all densities for our semiparametric model. Incorporating the only restriction on such densities of independence of (X_1, Y_1) and Z, it follows that $\mathcal{P}$ has elements of the form, in obvious notation, $p(v) = p(x_1, y_1)p(x_2|x_1, y_1, z)p(y_2|x_1, y_1, x_2, z)p(z|x_1, y_1)$, where $p(z|x_1, y_1) = \delta^z(1-\delta)^{1-z}$ and δ is known. The tangent space Γ is the mean square closure of the tangent spaces of parametric submodels

$$p(x_1, y_1; \xi_1)p(y_2|x_1, y_1, z; \xi_2) \cdot p(x_2|x_1, y_1, y_2, z; \xi_3)\delta^z(1-\delta)^{1-z}, \tag{A.4}$$

say. Each of the first three components of (A.4) must contain the truth. For example, if $p_0(x_2|x_1, y_1, y_2, z)$ is the true conditional density of X_2 given (X_1, Y_1, Y_2, Z), then, for h_3 such that $E\{h_3(X_1, Y_1, X_2, Y_2, Z)|X_1, Y_1, Y_2, Z\} = 0$, a typical submodel for this component is

$$p(x_2|x_1, y_1, y_2, z; \xi_3) = p_0(x_2|x_1, y_1, y_2, z) \cdot \{1 + \xi_3 h_3(x_1, y_1, x_2, y_2, z)\},$$

where ξ_3 is sufficiently small so that $p(x_2|x_1, y_1, y_2, z; \xi_3)$ is a density and the score with respect to ξ_3 may be shown to be $h_3(X_1, Y_1, Y_2, Z)$, and similarly for the first two components of (A.4). Evidently $\Gamma = \Gamma_1 \oplus \Gamma_2 \oplus \Gamma_3$, where (e.g., Newey, 1990)

$$\Gamma_1 = \{\text{all } h_1(X_1, Y_1) \in \mathcal{H}\} \quad [\text{so } E\{h_1(X_1, Y_1)\} = 0],$$

$$\Gamma_2 = [h_2(X_1, Y_1, Y_2, Z) \in \mathcal{H} \text{ such that } E\{h_2(X_1, Y_1, Y_2, Z)|X_1, Y_1, Z\} = 0],$$

$$\Gamma_3 = [h_3(X_1, Y_1, X_2, Y_2, Z) \in \mathcal{H} \text{ such that } E\{h_3(X_1, Y_1, X_2, Y_2, Z)|X_1, Y_1, Y_2, Z\} = 0].$$

It is easy to verify that Γ_1, Γ_2 and Γ_3 are all mutually orthogonal; e.g., for $h_2 \in \Gamma_2$, $h_3 \in \Gamma_3$,

$$E\{h_2(X_1, Y_1, Y_2, Z)h_3(X_1, Y_1, X_2, Y_2, Z)\} = E[h_2(X_1, Y_1, Y_2, Z) \cdot E\{h_3(X_1, Y_1, X_2, Y_2, Z)|X_1, Y_1, Y_2, Z\}] = 0.$$

Thus, $\Gamma^{\perp}$ is the space orthogonal to all of Γ_1, Γ_2 and Γ_3. It is straightforward to verify that the space $\Gamma_4 = [h_4(X_1, Y_1, Z) \in \mathcal{H}$ such that $E\{h_4(X_1, Y_1, Z)|X_1, Y_1\} = 0]$ is orthogonal to all of Γ_1, Γ_2 and Γ_3. Moreover, it may also be deduced that $\Gamma_1 \oplus \Gamma_2 \oplus \Gamma_3 \oplus \Gamma_4$ is in fact the entire Hilbert space $\mathcal{H}$ of mean-zero functions of V. Thus, it follows that Γ_4 contains all elements of $\mathcal{H}$ orthogonal to Γ, so that $\Gamma^{\perp} = \Gamma_4$.

Because Z is binary, we may write any element in $\Gamma^{\perp}$ equivalently as $Zh^{(1)}(X_1, Y_1) + (1-Z)h^{(0)}(X_1, Y_1)$ for some $h^{(c)}(X_1, Y_1)$, $c = 0, 1$, with finite variance such that $E\{Zh^{(1)}(X_1, Y_1) + (1-Z)h^{(0)}(X_1, Y_1)|X_1, Y_1\} = 0$. This implies $h^{(1)}(X_1, Y_1) = -h^{(0)}(X_1, Y_1) \cdot (1-\delta)/\delta$ for arbitrary $h^{(0)}(X_1, Y_1)$, showing that elements in $\Gamma^{\perp}$ may be written as $(Z-\delta)h(X_1, Y_1)$ for arbitrary h with $\operatorname{var}\{h(X_1, Y_1)\} < \infty$. Equivalently, we may write these elements as $-(Z-\delta)h(X_1, Y_1)$, which proves convenient in later arguments.

Recalling that $\mu_2^{(1)} = \mu_2 + \beta$ and $\mu_2^{(0)} = \mu_2$ and combining the foregoing results, we thus have that all influence functions for RAL estimators for β must be of the form

$$\frac{Z(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(1-Z)(Y_2 - \mu_2)}{1-\delta} - (Z-\delta)h(X_1, Y_1), \quad \operatorname{var}\{h(X_1, Y_1)\} < \infty, \tag{A.5}$$

which may also be expressed in the equivalent form given in (3).

We may in fact identify the efficient influence function φ^{eff} in class (A.5). By Result A.2 we may represent $\varphi^{\text{eff}}(X_1, Y_1, Y_2, Z) = \varphi^*(X_1, Y_1, Y_2, Z) - \Pi(\varphi^*|\Gamma^{\perp})$ for any arbitrary influence function φ^*, and, from above, we know that $\Pi(\varphi^*|\Gamma^{\perp})$ must be of the form $-(Z-\delta)h^{\text{eff}}(X_1, Y_1)$ for some h^{eff}. Projection is a linear operation; hence, taking φ^* to be (A.3), the projection may be found as the difference of the projections of each term in (A.3) separately. Moreover, by definition the residual for each term must be orthogonal to $\Gamma^{\perp}$. Thus, we wish to find $h^{\text{eff}\,(c)}(X_1, Y_1)$, $c = 0, 1$, such that

$$E\left(\left[\frac{Z(Y_2 - \mu_2^{(1)})}{\delta} - \left\{-(Z-\delta)h^{\text{eff}(1)}(X_1, Y_1)\right\}\right] \cdot (Z-\delta)h(X_1, Y_1)\right) = 0, \tag{A.6}$$

$$E\left(\left[\frac{(1-Z)(Y_2 - \mu_2^{(0)})}{1-\delta} - \left\{-(Z-\delta)h^{\text{eff}(0)}(X_1, Y_1)\right\}\right] \cdot (Z-\delta)h(X_1, Y_1)\right) = 0 \tag{A.7}$$

for all $h(X_1, Y_1)$. For (A.6), then, we require

$$E\left[\left\{\frac{Z(Y_2 - \mu_2^{(1)})}{\delta} + (Z-\delta)h^{\text{eff}(1)}(X_1, Y_1)\right\} \cdot (Z-\delta)\Big| X_1, Y_1\right] = 0 \quad \text{a.s.},$$

and similarly for (A.7). Using independence of (X_1, Y_1) and Z, we obtain

$$h^{\text{eff}(c)}(X_1, Y_1) = (-1)^c \frac{\{E(Y_2|X_1, Y_1, Z=c) - \mu_2^{(c)}\}}{\delta^c (1-\delta)^{1-c}}, \quad c = 0, 1.$$

For example, for $c = 1$ this follows from

$$\begin{aligned} &E\{Z(Z-\delta)(Y_2 - \mu_2^{(1)})|X_1, Y_1\} \\ &\quad = E[Z(Z-\delta)E\{(Y_2 - \mu_2^{(1)})|X_1, Y_1, Z\}|X_1, Y_1] \\ &\quad = (1-\delta)E\{(Y_2 - \mu_2^{(1)})|X_1, Y_1, Z=1\} \\ &\qquad \cdot P(Z=1|X_1, Y_1), \end{aligned}$$

where $P(Z=1|X_1, Y_1) = \delta$, and similarly

$$\begin{aligned} &E\{(Z-\delta)^2 h^{\text{eff}(1)}(X_1, Y_1)|X_1, Y_1\} \\ &\quad = \delta(1-\delta)h^{\text{eff}(1)}(X_1, Y_1). \end{aligned}$$

Substituting in $\varphi^*(X_1, Y_1, Y_2, Z) - \Pi(\varphi^*|\Gamma^{\perp})$, the efficient influence function is

$$\begin{aligned} &\left[\frac{Z(Y_2 - \mu_2 - \beta)}{\delta} - \frac{(Z-\delta)\{E(Y_2|X_1, Y_1, Z=1) - \mu_2 - \beta\}}{\delta}\right] \\ &\quad - \left[\frac{(1-Z)(Y_2 - \mu_2)}{1-\delta} + \frac{(Z-\delta)\{E(Y_2|X_1, Y_1, Z=0) - \mu_2\}}{1-\delta}\right]. \end{aligned}$$

A.3. REPRESENTATION OF OBSERVED-DATA INFLUENCE FUNCTIONS

Robins, Rotnitzky and Zhao (1994) derived the form of observed-data influence functions in (11) by adopting the geometric perspective on semiparametric models outlined in Appendix A.1. In contrast to the full-data situation of Appendix A.2, the relevant

Hilbert space $\mathcal{H}^{\text{obs}}$, say, in which observed-data influence functions are elements is now that of all mean-zero, finite-variance random functions $h(O)$, with analogous inner product and norm, that is, such functions depending on the observed data. The key is to identify the appropriate linear subspaces of $\mathcal{H}^{\text{obs}}$ (e.g., $\Gamma^{\text{obs}\perp}$ say) to deduce a representation of the influence functions, which in the general semiparametric model is a considerably more complex and delicate enterprise than for full-data problems.

We noted in Section 2.2 that, for purposes of deriving estimators for β based on the observed data, it suffices to identify observed-data influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$ separately. We now justify this claim. It is immediate from the definition (2) of an influence function that the differences of all observed-data influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$ are influence functions for observed-data estimators for β. Conversely, we may show that all observed-data influence functions for estimators for β can be written as the difference of observed-data influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$. In particular, if $\varphi_1(O)$ and $\varphi_0(O)$ are any observed-data influence functions for estimators for $\mu_2^{(1)}$ and $\mu_2^{(0)}$, respectively, then $\varphi_1(O) - \varphi_0(O)$ is an influence function for β by the above reasoning. By Result A.1 it follows that any observed-data influence function for an estimator for β can be written as $\varphi_1(O) - \varphi_0(O) + \psi(O)$, where $\psi(O) \in \Gamma^{\text{obs}\perp}$. We may rewrite this as $\{\varphi_1(O) + \psi(O)\} - \varphi_0(O)$. However, by Result A.1 $\{\varphi_1(O) + \psi(O)\}$ is an observed-data influence function for an estimator for $\mu_2^{(1)}$, concluding the argument.

A.4. DERIVATION OF THE EFFICIENT OBSERVED DATA INFLUENCE FUNCTION

Robins, Rotnitzky and Zhao (1994) provide a general mechanism for deducing the form of the efficient influence function. In the pretest–posttest problem this approach may be used to find the optimal choices for $h^{(1)}$ and $g^{(1)'}$ in (13) given in (14). However, because this mechanism is very general, for a simple model as in the pretest–posttest problem it is more direct and instructive to identify these choices via geometric arguments, as we now demonstrate.

We wish to determine $h^{\text{eff}(1)}$ and $g^{\text{eff}(1)'}$ such that the variance of (13) is minimized; that is, writing (13) as $A - B_1 - B_2$, as $E(A - B_1 - B_2) = 0$, we wish to minimize $E\{(A - B_1 - B_2)^2\}$. Geometrically, this is equivalent to finding the projection of A onto the subspace of $\mathcal{H}^{\text{obs}}$ of (mean-zero) functions of the form $B_1 + B_2$. It is straightforward to show that B_1 and B_2 are uncorrelated, whence it follows that, as $E\{(A - B_1 - B_2)^2\} = E\{(A - B_1)^2\} + E\{(A - B_2)^2\} - E(A^2)$ under these conditions, this minimization is equivalent to minimizing the variances of $A - B_1$ and $A - B_2$ separately. Because B_1 and B_2 are uncorrelated, they define orthogonal subspaces of $\mathcal{H}^{\text{obs}}$, so that these minimizations may be viewed as finding the separate projections of A onto these subspaces. Thus, as for the full-data case in Section A.2, we wish to find $h^{\text{eff}(1)}(X_1, Y_1)$ and $g^{\text{eff}(1)'}(X_1, Y_1, X_2, Z)$ such that, for all $h^{(1)}$ and $g^{(1)'}$,

$$E\Bigg(\Bigg[\frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi(X_1, Y_1, X_2, Z)} - \Bigg\{\frac{(Z-\delta)}{\delta} h^{\text{eff}(1)}(X_1, Y_1)\Bigg\}\Bigg] \cdot \frac{(Z-\delta)}{\delta} h^{(1)}(X_1, Y_1)\Bigg) = 0,$$

$$E\Bigg(\Bigg[\frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi(X_1, Y_1, X_2, Z)} - g^{\text{eff}(1)'}(X_1, Y_1, X_2, Z) \cdot \frac{\{R - \pi(X_1, Y_1, X_2, Z)\}}{\delta\pi(X_1, Y_1, X_2, Z)}\Bigg] \cdot g^{(1)'}(X_1, Y_1, X_2, Z) \cdot \frac{\{R - \pi(X_1, Y_1, X_2, Z)\}}{\delta\pi(X_1, Y_1, X_2, Z)}\Bigg) = 0.$$

A conditioning argument as in Section A.2 using $E(R|X_1, Y_1, X_2, Y_2, Z) = \pi(X_1, Y_1, X_2, Z)$ under MAR then leads to (14). In (14), $g^{\text{eff}(1)'}$ does not depend on $h^{\text{eff}(1)}$, and $h^{\text{eff}(1)}$ is identical to the optimal full-data choice in (4). These features *need not* hold for general semiparametric models; in particular, the choice of $\varphi^F(V)$ in (11) that yields the efficient observed-data influence function will *not* be the efficient full-data influence function in general. Here, this is a consequence of the simple pretest–posttest structure.

A.5. DEMONSTRATION OF (17)

The form of the influence function (17) when γ in (16) is estimated follows from a general result shown by Robins, Rotnitzky and Zhao (1994). In particular, Robins, Rotnitzky and Zhao showed precisely that, in our context, the influence function for the estimator for $\mu_2^{(1)}$ found by deriving an estimator for $\mu_2^{(1)}$ from the influence function $\psi(X_1, Y_1, X_2, R, RY_2, Z)$

in (15) (assuming $\pi^{(1)}$ is known) and then substituting an estimator for γ, where γ is estimated efficiently (e.g., by ML), is the residual from projection of $\psi(X_1, Y_1, X_2, R, RY_2, Z)$ onto the linear subspace of $\mathcal{H}^{\text{obs}}$ spanned by the score for γ. To demonstrate this, consider the special case of IWCC in (10), that is, (15) with $h^{(1)} \equiv q^{(1)} \equiv 0$. Suppose γ is estimated by ML from data with $Z = 1$ only. The score for γ is $S_\gamma(X_1, Y_1, X_2, Z; \gamma_0) = d(X_1, Y_1, X_2) \cdot \{R - \pi^{(1)}(X_1, Y_1, X_2; \gamma_0)\}Z$ and the relevant linear subspace of $\mathcal{H}^{\text{obs}}$ is $\{BS_\gamma(X_1, Y_1, X_2, Z; \gamma_0)$ for all $(p \times s)$ matrices $B\}$. Here, $b_{q(1)} = 0$ and the projection of ψ onto this space, $B_0 S_\gamma(X_1, Y_1, X_2, Z, \gamma_0)$, say, must satisfy

$$E\Bigg[\Bigg\{\frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi^{(1)}(X_1, Y_1, X_2; \gamma_0)} - B_0 S_\gamma(X_1, Y_1, X_2, Z, \gamma_0)\Bigg\} \cdot BS_\gamma(X_1, Y_1, X_2, Z, \gamma_0)\Bigg] = 0$$

for all B. By a conditioning argument similar to those in Appendices A.2 and A.4, we may find B_0 and show the projection is equal to the second term in the influence function

$$\frac{RZ(Y_2 - \mu_2^{(1)})}{\delta\pi^{(1)}(X_1, Y_1, X_2)} - d^T(X_1, Y_1, X_2)A_{(1)}^{-1}b_{(1)} \cdot \frac{\{R - \pi^{(1)}(X_1, Y_1, X_2)\}Z}{\delta} \tag{A.8}$$

and that (A.8) is (17) in this special case.

As noted in Section 3.5, for choices of $h^{(1)}$ and $q^{(1)}$ other than the optimal ones, estimating γ even if it is known leads to a gain in efficiency. Geometrically this is because (17) is the residual found from projection of ψ onto a linear subspace of $\mathcal{H}^{\text{obs}}$.

A.6. DEMONSTRATION OF DOUBLE ROBUSTNESS PROPERTY 2

We must show that the right-hand side of (18) converges in probability to $\mu_2^{(1)}$ if the true $\pi^{(1)}$ is replaced by an incorrect model π^*. Multiplying and dividing each term by n and using $n_1/n \to \delta$, that the second term converges in probability to zero is immediate by the independence of Z and (X_1, Y_1). The first term converges in probability to

$$E\left\{\frac{RZY_2}{\delta\pi^*(X_1, Y_1, X_2)}\right\} = E\left\{\frac{Z\pi^{(1)}(X_1, Y_1, X_2)}{\delta\pi^*(X_1, Y_1, X_2)}Y_2\right\} = E\left\{\frac{Z\pi^{(1)}(X_1, Y_1, X_2)}{\delta\pi^*(X_1, Y_1, X_2)}E(Y_2|X_1, Y_1, X_2, Z)\right\}$$

by a conditioning argument similar to those above. Similarly, the third term converges to

$$E\Bigg[\frac{Z\{\pi^{(1)}(X_1, Y_1, X_2) - \pi^*(X_1, Y_1, X_2)\}}{\delta\pi^*(X_1, Y_1, X_2)} \cdot E(Y_2|X_1, Y_1, X_2, Z)\Bigg],$$

using $ZE(Y_2|X_1, Y_1, X_2, Z = 1) = ZE(Y_2|X_1, Y_1, X_2, Z)$. Thus, their difference converges to $E\{E(ZY_2|X_1, Y_1, X_2, Z)\}/\delta = E\{ZE(Y_2|Z)\}/\delta = E(Y_2|Z = 1)$ as in Section 3.2.

ACKNOWLEDGMENTS

The authors are grateful to Michael Hughes, Heather Gorski and the AIDS Clinical Trials Group for providing the ACTG 175 data. This research was supported in part by Grants R01-CA051962, R01-CA085848 and R37-AI031789 from the National Institutes of Health.

REFERENCES

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BROGAN, D. R. and KUTNER, M. H. (1980). Comparative analyses of pretest–posttest research designs. *Amer. Statist.* **34** 229–232.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.

CLEVELAND, W. S., GROSSE, E. and SHYU, W. M. (1993). Local regression models. In *Statistical Models in S* (J. M. Chambers and T. J. Hastie, eds.) 309–376. Wadsworth, Pacific Grove, CA.

CRAGER, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baseline. *Biometrics* **43** 895–901.

FOLLMANN, D. A. (1991). The effect of screening on some pretest–posttest test variances. *Biometrics* **47** 763–771.

HAMMER, S. M., KATZENSTEIN, D. A., HUGHES, M. D., GUNDAKER, H., SCHOOLEY, R. T., HAUBRICH, R. H., HENRY, W. K., LEDERMAN, M. M., PHAIR, J. P., NIU, M., HIRSCH, M. S. and MERIGAN, T. C., for The AIDS Clinical Trials Group Study 175 Study Team (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England J. Medicine* **335** 1081–1090.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685.

KOCH, G. G., TANGEN, C. M., JUNG, J.-W. and AMARA, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17** 1863–1892.

LAIRD, N. (1983). Further comparative analyses of pretest–posttest research designs. *Amer. Statist.* **37** 329–330.

LEON, S., TSIATIS, A. A. and DAVIDIAN, M. (2003). Semiparametric estimation of treatment effect in a pretest–posttest study. *Biometrics* **59** 1046–1055.

LUENBERGER, D. G. (1969). *Optimization by Vector Space Methods*. Wiley, New York.

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* **23** 2937–2960.

NEWEY, W. K. (1990). Semiparametric efficiency bounds. *J. Applied Econometrics* **5** 99–135.

ROBINS, J. M. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In *ASA Proc. Bayesian Statistical Science Section* 6–10. Amer. Statist. Assoc., Alexandria, VA.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866.

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

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Rejoinder to "Adjusting for nonignorable drop-out using semiparametric nonresponse models." *J. Amer. Statist. Assoc.* **94** 1135–1146.

SINGER, J. M. and ANDRADE, D. F. (1997). Regression models for the analysis of pretest/posttest data. *Biometrics* **53** 729–735.

STANEK, E. J., III (1988). Choosing a pretest–posttest analysis. *Amer. Statist.* **42** 178–183.

STEIN, R. A. (1989). Adjusting treatment effects for baseline and other predictor variables. In *ASA Proc. Biopharmaceutical Section* 274–280. Amer. Statist. Assoc., Alexandria, VA.

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York.

YANG, L. and TSIATIS, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *Amer. Statist.* **55** 314–321.

Comment

Hyonggin An and Roderick Little

We congratulate the authors on a very useful article. The semiparametric approaches of Robins and coworkers have attracted considerable attention among theoretically-inclined statisticians, but it appears to us that the difficulties of understanding exactly how to implement the approach in real problems has deterred many practitioners from applying the methods. This application of the methods to a common issue in biostatistical analysis is thus most welcome, and we are pleased to have the opportunity to comment.

Our attitude to the methodology mirrors the situation of a consumer at an electronics store who is trying to keep apace with the advances in electronic wizardry. We have a PSM (Predictive Statistical Modeling) machine that we like, which has a flexible set of options, some of which we have figured out how to use, and others which have not attracted our attention sufficiently for us to try to master them. The Robins Company has now developed a new spiffy SPDRWEE (Semi-Parametric Doubly Robust Weighted Estimating Equation) model, which beguiles us with offers of increased power and flexibility. The only problem is that the instruction manual is even more complicated than the one for our current model, which we have just begun to master, and the dials and switches on the new model are located in different places. Our question (particular from the author whose capacity to absorb new ideas has been regrettably tarnished by age) is whether the new model is a major breakthrough, or whether we can continue to live with the current model.

Our PSM machine's approach to the pretest–posttest trial without missing data is to regress the outcome on the treatment dummy and baseline covariates that are predictive of the outcome. Parametric modeling of the baseline covariates should suffice, since the randomization ensures balance with the respect to these

Hyonggin An is Assistant Professor, Department of Biostatistics, University of Iowa, 200 Hawkins Drive, Iowa City, Iowa 52242-1009, USA (e-mail: hyonggin-an@uiowa.edu). Roderick Little is Richard D. Remington Collegiate Professor of Biostatistics, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA (e-mail: rlittle@umich.edu).

variables, so misspecification of the parametric model may slightly reduce efficiency but will not lead to bias. In principle we believe that baseline variables should be adjusted since they are causally prior to the treatment; the question of whether to adjust for pretest values and/or their interactions with treatment dummies is a model selection issue. The authors state that "no general consensus has emerged regarding the preferred approach," but the methods covered in this comment all fall under the general rubric of regression modeling, and there can be no "general consensus" any more than one can make general rules about which variables to include in any real-world regression. We see no need to change our existing modeling philosophy to address these variable selection issues, since they are not answered by the Robins et al. theory any more than in the PSM framework.

We now turn to the problem of missing data in the outcome variable. With no intervening variables, and assuming MAR, the incomplete cases carry no information for the regression and can be discarded (see, e.g., Little and Rubin, 2002, Section 2.3). The regression modeling of the resulting complete cases needs more attention to parametric assumptions, to the extent that the balance from randomization has been disturbed, but the problem is still essentially one of regression, so we see nothing very new here. (Two minor quibbles: the authors' "paired t-test" method is actually an unpaired t test on the posttest–pretest differences; and contrary to the authors' statement, complete-case regression methods allow missingness to depend on the covariates included as predictors, and hence do not assume MCAR.)

With intervening variables, the incomplete cases potentially carry information about the missing outcomes, but including them as predictors in the regression model is inappropriate since they are post-treatment variables. To illustrate, we ran a regression of the outcome on the complete cases that included the covariates and intermediate variables (namely, wt, HIV, prior, Karn, $CD8_0$, ${CD8_0}^2$, $CD4_0$, ${CD4_0}^2$, $CD82_0$, ${CD82_0}^2$, $CD42_0$, ${CD42_0}^2$, offtrt and treatment). This yielded a treatment effect for the data in question of 31.14 (SE 7.93), which is quite a bit lower than other estimates, and we expect is biased downwards by the inappropriate adjustment for the intermediate variables.

The question, then, is how to include the information in the incomplete cases without adjusting for the intermediate variables in the final regression. If the latter are earlier measures of the outcome, such as intermediate CD4, then this can be achieved by a parametric linear mixed model, fitted by maximum likelihood—the authors' claim that this approach is not widely used by practitioners surprises us, since our experience is that it is widely used in the pretest–posttest context. As with any model there are dangers of misspecification of the form of the regression model, but mixed models allow random effects to model the association between the repeated measures over time without adjusting them in the analysis. We do not pursue this approach in detail here, since there are intermediate variables other than CD4 that are not naturally modeled by a repeated-measures approach.

A parametric PSM approach to the problem is to impute the outcomes based on the baseline and intermediate variables, and compute the regression of the outcome on the treatment dummy and baseline covariates using the filled-in data. A general approach is to apply multiple imputation (MI), with imputation uncertainty assessed by Rubin's MI combining rules; in this simple setting we simply imputed conditional means and assessed uncertainty by bootstrapping the whole procedure (e.g., Little and Rubin, 2002, Chapter 5). The imputation step can be achieved using a normal model, or more flexible MI methods such as the sequential imputation algorithms in IVEWARE (Raghunathan et al., 2001) or MICE (van Buuren and Oudshoorn, 1999). This analysis takes advantage of an attractive feature of MI, namely that the imputation model does not have to be the same as the analysis model. Here the imputation model conditions on the intermediate variables, but the final regression model does not.

This approach relies on a correct specification of the imputation model, particularly if there are a lot of missing data—the analysis model is protected by the balance from the randomization. What to do if we want to avoid such parametric assumptions? With a single covariate one might base imputations on a smooth nonparametric function of the covariate, such as a kernel regression model (e.g., Cheng, 1994). However, with multiple predictors nonparametric specification of the imputation model is subject to the so-called "curse of dimensionality," which inhibits the ability to fit spline-like regression models without assumptions of additivity, as are made in generalized additive models.

We recently proposed a semiparametric approach that addresses the "curse of dimensionality" within the PSM framework, which we call propensity spline prediction (Little and An, 2004). In this approach, the propensity to be missing is modeled by a logistic regression of the missing-data indicator on the baseline and intermediate variables. The missing outcomes are then imputed by a penalized spline on the estimated

propensity (Eilers and Marx, 1996). Other predictors are also regressed on penalized splines of the response propensity, and the residuals from these regressions are added parametrically to the imputation model. The key idea is that the relationship between the outcome and the propensity to respond needs to be modeled correctly to avoid bias, and hence an estimate of the propensity to respond is included nonparametrically in the prediction model. The other variables can be added parametrically to increase precision; since respondents and nonrespondents are balanced with respect to these variables conditional on the propensity score, misspecification of the parametric form does not result in bias. Little and An (2004) discuss a "double robustness property" for this method, and claim that propensity spline prediction obviates the need for the "calibration" correction in the SPDRWEE approach.

In the application described, the PSM approach with a parametric imputation model yields an estimated treatment effect of 52.76 (Bootstrap SE = 11.64), and the propensity spline prediction approach yields an estimated treatment effect of 52.37 (Bootstrap SE = 10.92). These estimates are not very different from the IWCC method, which yields a treatment effect of 54.69 (SE = 11.79), or the SPDRWEE method, which yields an estimate of 57.24 (SE = 10.20). The question of which method is better cannot be answered by comparing results for a single data set, but our prediction approaches are more comparable to the authors' methods than the ANCOVA and paired t-test methods in the authors' Table 1, since they make full use of the observed information and capitalize on the observed covariates.

We suggest that the existing principles of regression modeling, made more robust by the propensity spline prediction, provide good answers for the problem described, without the need for a new "unified framework" of inference.

Comment

Babette A. Brumback and Lyndia C. Brumback

1. INTRODUCTION

We are grateful to the Editor, George Casella, for the opportunity to discuss this elucidating paper by Davidian, Tsiatis and Leon. It is the first time either of us has seen such a concrete and accessible account of the semiparametric efficiency results of Robins and colleagues. The focus on the pretest–posttest problem with MAR posttest data coincides with a problem we have met several times in practice. In studying the ideas presented, we have spent much time scrutinizing the semiparametric efficient estimator (SPEE) given by the authors' equation (18) and paying special attention to its reductions under some commonplace simplifying restrictions. One of the most conspicuous features of the SPEE, under any circumstances, is an absence of dependence on the conditional variance of the posttest data. This prompts us to compare the SPEE to shrinkage estimators (Lehmann and Casella, 1998) that not only rely on conditional means of the posttest data, but also make use of conditional variances. To gradually build a better understanding of the SPEE, in our Section 2 we restrict attention to data only on a pretest score, a continuous posttest score and treatment assignment. In Section 3 we consider data only on an intermediate test score, a continuous posttest score and treatment assignment. In Section 4 we discuss dimension reduction via the probability of missingness and apply it in conjunction with a shrinkage estimator to reanalyze the ACTG 175 data. We conclude with a summary of the questions we have raised and answered, as well as some additional unanswered questions for the authors.

Babette A. Brumback is Associate Professor, Division of Biostatistics, Department of Health Services Research, Management, and Policy, University of Florida, Gainesville, Florida 32611, USA (e-mail: bbrumback@phhp.ufl.edu). Lyndia C. Brumback is Research Assistant Professor, Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232, USA (e-mail: lynb@u.washington.edu).

2. AUXILIARY DATA ON Y_1 ONLY

We first focus on the SPEE under the restriction that the posttest data Y_2 are MAR conditional on pretest data Y_1 and treatment assignment Z, and that data on additional baseline covariates X_1 and intermediate variables X_2 are irrelevant.

2.1 Connection with Lord's Paradox

Were one to discard data from participants missing Y_2, the statistical model for the remainder would be structurally identical to that for causal inference models with ignorable (Stone, 1993) treatment assignment conditional on Y_1. Because the missingness mechanism depends only on Y_1 and Z, and not on posttreatment variables, the complete-case data could just as well have been generated by randomizing treatment assignment Z within (possibly infinitesimal) strata defined by Y_1. The latter scenario was contemplated by Lord (1967), who identified a paradox that would sometimes occur if, perchance, statistician A contrasted the distribution of the change score $Y_2 - Y_1$ across Z, whereas statistician B contrasted the conditional distribution of the posttest score given the pretest score, $Y_2|Y_1$, across Z. Consider an idealized example in which $E(Y_2 - Y_1|Z = 1) - E(Y_2 - Y_1|Z = 0) = 0$ but $E(Y_2|Y_1, Z = 1) - E(Y_2|Y_1, Z = 0) = \beta > 0$. In this case, statistician A would conclude no difference, while statistician B would conclude a positive effect β of treatment. The paradox resolves when one recognizes that Y_1 not only confounds the effect of Z on Y_2, but also confounds the effect of Z on the change score $Y_2 - Y_1$. Thus Lord's paradox is really just an instance of Simpson's paradox, with $E(Y_2 - Y_1|Z = 1) - E(Y_2 - Y_1|Z = 0) = 0$ but $E(Y_2 - Y_1|Y_1, Z = 1) - E(Y_2 - Y_1|Y_1, Z = 0) = \beta > 0$.

The IWCC estimator would make use of essentially the same data available to statisticians A and B, and would consistently estimate β. The IWCC estimator relies on a correct model for either $P(R = 1|Y_1, Z)$ or $P(Z = 1|Y_1, R = 1)$, but does not need to model $E(Y_2|Y_1, Z)$. The SPEE, on the other hand, uses models for $E(Y_2|Y_1, Z)$. Suppose these are correctly modeled with $E(Y_2|Y_1, Z = 1) = \beta_1 + \alpha Y_1$ and $E(Y_2|Y_1, Z = 0) = \beta_0 + \alpha Y_1$, so that the true effect is $\beta \equiv \beta_1 - \beta_0$. Based on these models, the SPEE reduces to

$$\hat{\beta}_{\text{initial}} + \hat{E}\big[(Y_2 - \beta_1 - \alpha Y_1)R/\pi^{(1)}|Z = 1\big] - \hat{E}\big[(Y_2 - \beta_0 - \alpha Y_1)R/\pi^{(0)}|Z = 0\big].$$

The SPEE makes use of additional data from persons with missing Y_2 in the computation of $\hat{\beta}_{\text{initial}}$, which technically equals

$$(1) \quad (1/n)\sum_{\text{all } i}\big(\hat{E}(Y_2|Y_1, Z = 1) - \hat{E}(Y_2|Y_1, Z = 0)\big),$$

where the sum is over all individuals, including those with missing Y_2. We observe with this example how the various modeling choices for $E(Y_2|Y_1, Z)$ can lead to confusion as to the exact form of the SPEE: in working with (1), it is unclear whether we should handle the regressions $E(Y_2|Y_1, Z = 1)$ and $E(Y_2|Y_1, Z = 0)$ separately or first reduce their subtraction to β and then estimate in any way we wish.

2.2 Comparison with Shrinkage Estimators

We next assume that no posttest data are missing and that Y_1 can be dichotomized into $Y_1 = 0$ or $Y_1 = 1$ without loss of information for estimating β. Under these conditions, the SPEE for $\mu_2^{(1)}$ simply equals the unweighted mean of Y_2 in the treatment group, that is,

$$(1/n_1)\sum_i Z_i Y_{2i}.$$

Similarly the SPEE for $\mu_2^{(0)}$ equals the unweighted mean of Y_2 in the control group.

Although the SPEE is efficient in many circumstances, it is not always. As we next demonstrate, some scenarios give way to a preference for estimators that take into account the conditional variance of Y_2, for example, shrinkage estimators, which shrink imprecise cluster means toward precise cluster means when the means are close relative to their variability. For simplicity and without loss of generality, we focus during the remainder of this subsection on estimation of $\mu_2^{(1)}$. Rather than using the SPEE, which assigns weight m_j/n_1, $j = 0, 1$, $m_0 \equiv \sum_i Z_i(1 - Y_{1i})$, $m_1 \equiv \sum_i Z_i Y_{1i}$, to each cluster mean, that is, to

$$\bar{Y}_{20} = \left(\sum_i Z_i(1 - Y_{1i})Y_{2i}\right)\Big/ m_0$$

and

$$\bar{Y}_{21} = \left(\sum_i Z_i Y_{1i} Y_{2i}\right)\Big/ m_1$$

in the calculation of $\mu_2^{(1)}$, one might try instead a weighted average of the two cluster means with weights proportional to their inverse variances. Preferably, a compromise will be sought and determined by the distance between the two cluster means relative to their variances.

Specifically, we propose the compromise estimator (CE) for $\mu_2^{(1)}$ which can be derived under the mixed effects model

$$Y_2 = X\mu_2^{(1)} + Uu + \varepsilon,$$

with X the vector of ones (1_{n_1}), $\mu_2^{(1)}$ a fixed effect, U a two-column matrix with first column containing indicators $(1 - Y_1)$ and second column containing indicators Y_1, u a vector of two random effects independent of one another and each $N(0, \tau^2/w_j)$,

TABLE 1
Comparison of CE and SPEE (of $\mu_2^{(1)}$) in terms of mean squared error ($m_0 = m_1 = n_1/2$, $\sigma_0^2 = \sigma_1^2/2$)

$\bar{Y}_{20}$	$\bar{Y}_{21}$	τ^2	σ_1^2	n_1	CE	SPEE	MSE_{CE}	MSE_{SPEE}
0	0	0	1	1000	0	0	0.000667	0.00075
0	0.01	2.5e − 05	1	1000	0.0034	0.005	0.000669	0.00075
0	0.1	0.0025	1	1000	0.046	0.05	0.00073	0.00075
0	1	0.25	1	1000	0.5	0.5	0.00075	0.00075

$w_j = m_j/n_1$, $j = 0, 1$ and ε the vector of error terms, assumed independent of one another and of the random effects, with distribution $N(0, \sigma_0^2)$ for observations with $Y_1 = 0$ and $N(0, \sigma_1^2)$ for observations with $Y_1 = 1$.

Straightforward calculation shows that the CE of $\mu_2^{(1)}$ equals the weighted average of cluster means with weights proportional to

$$\frac{1}{\tau^2/w_0 + \sigma_0^2/m_0}$$

for cluster $Y_1 = 0$ and to

$$\frac{1}{\tau^2/w_1 + \sigma_1^2/m_1}$$

for cluster $Y_1 = 1$. If the two cluster means are equal, then $\tau^2 = 0$ and the CE weights proportionally to the inverse variance of each cluster mean. For example, when $\sigma_0^2 = \sigma_1^2/2$, $m_0 = m_1 = n_1/2$ and $\tau^2 = 0$, we weight observations with $Y_1 = 0$ twice as much as those with $Y_1 = 1$. However, when the cluster means are far apart, τ^2/w_j is large relative to σ_j^2/m_j, and the CE weights proportionally to w_j, exactly as the SPEE would do. For situations in between, the CE compromises between the two estimators based on the size of τ^2/w_j (which measures the distance between the two cluster means) and the σ_j^2/m_j (the variance of each cluster mean).

Why is the CE sometimes preferable to the SPEE? It allows for a smaller mean squared error under some circumstances, even when normality is not assumed, and a nearly equivalent mean squared error under other circumstances. Continuing with our example, in which $m_0 = m_1 = n_1/2$ and $\sigma_0^2 = \sigma_1^2/2$, when $\tau^2 = 0$ the mean squared error of the SPEE is $(3/4)\sigma_1^2/n_1$, while that of the CE is only $(2/3)\sigma_1^2/n_1$, and both estimators are unbiased. When $\tau^2 > 0$ the CE is biased in small samples, but its mean squared error (MSE) is less than that of the unbiased SPEE, until the sample size is large enough that the CE almost equals the SPEE (see Table 1). Because the univariate mean is admissible (Lehmann and Casella, 1998), there must be a region for τ^2 in which the unconditional MSE of the SPEE is less than that of the CE, but the table does not show it. This is because we approximated the MSE by not accounting for the estimation of τ^2. We leave as a conjecture for future study that the CE is itself admissible. Note that for Table 1 we estimated τ^2 as the variance of $(\bar{Y}_{20}\sqrt{w_0}, \bar{Y}_{21}\sqrt{w_1})$.

What is the relevance of this discussion for the general case with missing posttest data? We again find that the SPEE will tend toward an unweighted average of the weighted individual observations $Y_{2i}/\pi_i^{(1)}$, whereas the CE will compromise based on the conditional variance so as to reduce the mean squared error when the conditional mean is constant. Which estimator should we prefer in practice? This is a difficult question, mostly because in practice the conditional variances as well as the conditional means must be estimated, often leading to great uncertainty associated with either choice.

3. AUXILIARY DATA ON X_2 ONLY

In this section we shift focus to the case of posttest data Y_2 MAR conditional on intermediate data X_2 and treatment assignment Z, with pretest data Y_1 and additional baseline covariates X_1 irrelevant.

It is generally well known that conditioning on a variable that is affected by treatment and then subsequently affects the posttest can induce bias, typically by canceling out the indirect effect. However, if the missing data depend on an intermediate variable, we can neither ignore it in the analysis nor treat it identically as a pretreatment variable. How does SPEE recognize an intermediate variable from a baseline variable? The difference is encoded in the modeling assumption that $Z \amalg (Y_1, X_1)$ and the absence of the assumption that $Z \amalg X_2$. We also observe that X_2 enters the estimating equation only in the case of missing posttest data; otherwise, the third term of the SPEE equals zero.

We note that if $Y_2 = X_2$, the SPEE of $\mu_2^{(1)}$ equals $(1/n_1)\sum_i Z_i X_{2i}$. If instead $Y_2 = X_2 + \varepsilon$, the estimator becomes

$$\frac{1}{n_1}\sum_i Z_i R_i \frac{(\varepsilon_i)}{\pi^{(1)}(X_{2i})} + \frac{1}{n_1}\sum_i Z_i X_{2i}.$$

Thus, if Y_{2i} is observed, we upweight its residual, effectively making multiple copies, and we add one copy to the corresponding X_{2i} and the others to X_{2j}, corresponding to missing Y_{2j}. Then we average. The overall effect is to impute missing values Y_{2j} based on X_{2j} and one of the observed ε_i.

In general, we can make use of the MAR assumption to unbiasedly estimate $\mu_2^{(1)}$ via the equation

$$E[Y_2|Z=1] = E_{X_2|Z=1}E[Y_2|X_2, Z=1, R=1]. \quad (2)$$

It is important to recognize that the outer expectation is taken with respect to the conditional distribution of $X_2|Z=1$ rather than $X_2|Z=1, R=1$. The MAR assumption allows us to condition on $R=1$ in the inner expectation, because R is independent of Y_2 given X_2 and Z. We use the inner expectation to predict Y_2 given X_2 and $Z=1$ based on the observed data, and then we average over these predictions based on the distribution of $X_2|Z=1$ in the complete data set. In this procedure we both condition on X_2 and then uncondition on X_2, but in a way that does not leave us back with the obviously flawed estimator $\hat{E}[Y_2|Z=1, R=1]$.

4. COMPARATIVE ANALYSIS OF ACTG 175

4.1 Dimension Reduction via the Probability of Missingness

A consequence of assuming a known model for $\pi(X_1, Y_1, X_2, Z)$ is that for each participant, the multivariate data (X_1, Y_1, X_2, Z) can be reduced to the univariate data $Q \equiv \pi(X_1, Y_1, X_2, Z)$ when imputing the missing posttest scores. That is, rather than imputing Y_2 with a high-dimensional model for $E(Y_2|X_1, Y_1, X_2, Z)$ and $\mathrm{Var}(Y_2|X_1, Y_1, X_2, Z)$, we can instead impute based on a simpler model for $E(Y_2|Q, Z)$ and $\mathrm{Var}(Y_2|Q, Z)$. The proof is straightforward: that Y_2 is MAR conditional on (X_1, Y_1, X_2, Z) and that $\pi(X_1, Y_1, X_2, Z)$ is a known function implies that Y_2 is MAR conditional on Q and A, for A any function of (X_1, Y_1, X_2, Z).

4.2 An Alternative Methodology and its Application to ACTG 175

We next combine the dimension reduction based on Q, the shrinkage methodology outlined in Section 2.2, and a generalization of (2) to reanalyze the ACTG 175 data.

By the argument in Section 4.1, the MAR assumption allows us to estimate $E[Y_2|Q, Z=1]$ using the quantity $E[Y_2|Q, Z=1, R=1]$ based on observed data only. Thus, we find that we can estimate $\mu_2^{(1)}$ via

$$E[Y_2|Z=1] = E_{Q|Z=1}E[Y_2|Q, Z=1, R=1], \quad (3)$$

similarly to (2). That is, we first regress Y_2 on Q using complete case data in the treated group and then we average the predictions based on this model using the distribution of Q on everyone in the treated group, including those with missing Y_2. This gives us an unbiased estimator of $\mu_2^{(1)}$ that is easy to compute, but that is not necessarily efficient for two reasons. The first is that by reducing the data via Q we lose the ability to use the rest of X_1, Y_1, X_2 for efficiency purposes. The second is that using shrinkage estimators for $E[Y_2|Q, Z=1, R=1]$ or in the averaging of that quantity with respect to $E_{Q|Z=1}$ can lead to efficiency gains, as in Section 2.2.

It is computationally more difficult but still theoretically feasible to increase efficiency either by using more than Q in the estimation of $\mu_2^{(1)}$, that is, by basing estimation on

$$E[Y_2|Z=1] = E_{(Q,A)|Z=1}E[Y_2|Q, A, Z=1, R=1] \quad (4)$$

rather than on (3), or by using shrinkage ideas that compromise between averaging with respect to $Q|Z=1$ (to produce an unbiased estimator) and averaging with respect to the inverse of $\mathrm{Var}[Y_2|Q, Z=1, R=1]$ (to produce an estimator that would be efficient and unbiased if $E[Y_2|Q, Z=1, R=1]$ did not depend on Q). One could also combine the two approaches, trading between an average based on $(Q, A)|Z=1$ and one based on the inverse of $\mathrm{Var}[Y_2|Q, A, Z=1, R=1]$.

For expository purposes, we reanalyze the ACTG 175 data based on Q only (i.e., letting A be empty). We first estimate $\pi(X_1, Y_1, X_2, Z)$ exactly as did the authors, to obtain Q. We then focus on estimating $E[Y_2|Q, Z, R=1]$ within each treatment group separately. The scatterplots in Figure 1 show Y_2 versus Q within each treatment group, and the rug plots detail

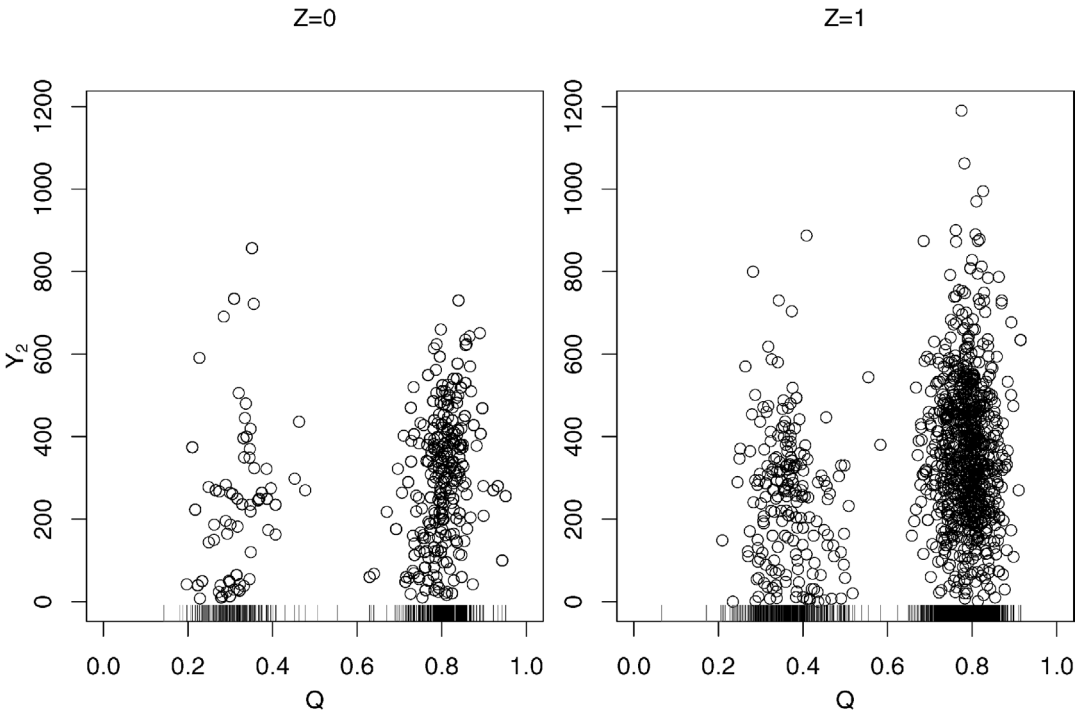

FIG. 1. *Y_2 versus Q by Z.*

the distribution of $Q|Z$ for all individuals (and not just those with $R = 1$). Interestingly, for each treatment group the scatterplot separates into two distinct clusters. It turns out that the contribution of the *off-treatment* variable overwhelms that of the other variables in the calculation of Q. When we dichotomize Q into $Q \equiv 1$ (original $Q \geq 0.6$), Q is identical to one minus *off-treatment* for all but five people. For ease of illustration, we use the dichotomized version of Q. This leads to the statistical summary in Table 2.

We first compute the estimator (3), which weights each cluster mean according to the total number of observations, as follows. For $Z = 0$, the $Q = 0$ cluster mean is weighted by $213/532 = 0.40$ and the $Q = 1$ mean is weighted by 0.60. Thus (3) estimates $\mu_2^{(0)}$ at $241.8 \times 0.4 + 299.5 \times 0.6 = 276.4$. For $Z = 1$, the $Q = 0$ cluster mean is weighted by $562/1607 = 0.35$ and the $Q = 1$ mean is weighted by 0.65. This estimates $\mu_2^{(1)}$ at 324.7. The estimate of β is thus 48.3. We approximate the standard error as

$$\{23.8^2 \times 0.4^2 + 9.8^2 \times 0.6^2 + 11.5^2 \times 0.35^2 + 6.0^2 \times 0.65^2\}^{1/2} = 12.5.$$

TABLE 2

	Control Group		**Treatment Group**	
	$Q = 0$	$Q = 1$	$Q = 0$	$Q = 1$
Mean of Y_2	241.8	299.5	253.9	362.7
SE of mean of Y_2	23.8	9.8	11.5	6.0
Number nonmissing Y_2	66	255	200	821
Total number	213	319	562	1045

To calculate the CE, instead of weighting by, for example, 0.4 and 0.6 in the $Z = 0$ group, we estimate τ^2 and weight proportionally to $1/(\tau^2/0.4 + 23.8^2)$ and $1/(\tau^2/0.6 + 9.8^2)$. We estimate τ^2 as 3126. Since τ^2/w_j is so much larger than σ_j^2/m_j in each cluster, the CE is effectively identical to (3).

It is interesting to compare the CE, which equals 48.3 with a standard error of 12.5, to the authors' SPEE, which equals 57.24 with a standard error of 10.2, and the authors' ANCOVA, which equals 64.54 with a standard error of 9.33. The ANCOVA estimator uses data on complete cases only. Perhaps because these participants tend to have $Q = 1$, the ANCOVA estimator is close to the estimator obtained by incorrectly stratifying on $Q = 1$, which gives $362.7 - 299.5 = 63.2$. The CE seems to incorporate more information from the $Q = 0$ group than does the SPEE. Perhaps if we were to treat Q continuously, we would find an estimator closer to the SPEE, or perhaps we need to incorporate additional covariates, rather than letting A be empty. Because the variance of Y_2 seems to depend on Y_1 in

the authors' figure, it would be interesting to do further analyses letting $A \equiv Y_1$. This would raise many practical issues for calculating the CE; for instance, with three clusters rather than two, should one shrink two of the three cluster means together if they are "close" relative to their variances, or would one only shrink in the scenario that all three cluster means are close to one another? With a continuous A and/or Q, we move from a discrete number of clusters to the continuous setting in which a finite basis must be selected. The practical issues become even more complex.

5. SUMMARY AND ADDITIONAL QUESTIONS

Our discussion has explored the SPEE under some commonplace simplifying restrictions. We have also compared the SPEE to shrinkage estimators based on a dimension reduction via the probability of missingness. We briefly discussed the incorporation of other covariates via (4). This leaves as a question for future research how to select A to calculate efficient estimators, which might be considered parallel to the problem of estimating the unknown conditional expectations during computation of the SPEE. Furthermore, there remain several practical issues alluded to at the end of Section 4 that involve shrinkage when Q and/or A are continuous.

We conclude our discussion with three additional questions for the authors. First, given that many of the participants in ACTG 175 go off treatment, why has an intent-to-treat parameter been chosen as the target of inference? This is most likely for ease of illustration, but it would be of further interest to apply methods for noncompliance [e.g., as discussed by Robins (1994)] to assess treatment efficacy rather than programmatic effectiveness.

Second, how would the methodology of Robins and colleagues unfold if the class of influence functions were narrowed to include only those with $E(\phi(W)|V) = 0$ for V some subset of W? For example, in the authors' analysis V could be Y_1. How would the efficient estimator for the class of $E(\phi(W)) = 0$ relate to that for the class of $E(\phi(W)|V) = 0$? Additionally, were we to narrow the class of influence functions in this way, would the authors then recommend conducting conditional (on V) or unconditional inference?

Third, we wonder about the practical issues associated with model choice in computation of the SPEE. The authors comment in Section 5 that by basically using larger component models (in calculating the conditional expectations), one will obtain more efficient results. Why, mathematically, might this be so? Surely there must be a breakdown of this phenomenon in practical sample sizes. Related to this, could the SPEE be derived via our (4)? Equation (4) produces robust estimators: when Q is misspecified but the equation still holds, we achieve consistency. Also, could basing estimation on (4) lead to straightforward transfer of model choice procedures designed for standard regression?

Comment

Geert Molenberghs

1. INTRODUCTION

The existing research area of incomplete data methodology is characterized by three main, interrelated issues. First, biopharmaceutical and other practice still sticks to amazingly simplistic and generally incorrect methods. Second and related, more advanced methodology, such as methods valid when data are missing at random and missing not at random, are perceived to be complicated and lacking unification. Third and equally related, the academic research community is divided between two rather opposing schools: the likelihood-oriented school of Rubin and co-workers, on the one hand, and the weighting-based school of Robins, Rotnitzky and co-workers, on the other hand. Exchanges between these two school can certainly be entertaining, but when debates are too fierce and go on for too long, the winner is likely to be a third party. In this case, the third party may well be *last observation carried forward* (LOCF), *complete case analysis* (CC) and related simplistic methods.

The tremendous merit of this paper is that it addresses these problems in a very successful way, using

Geert Molenberghs is Professor of Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium (e-mail: geert.molenberghs@luc.ac.be).

sound yet accessible methodology. The authors steer clear of controversies merely rooted in principle and, instead, present a unifying framework. The use of the pretest–posttest setting, which in the authors' concept of the term encompasses almost every longitudinal setting (and, as such, the term may sound more limited than it is supposed to be), allows clear and understandable illustration of the concepts developed. The paper nicely illustrates how simple methods, such as two-sample, paired t tests and analysis of covariance, can be correct but inefficient with complete data, but in addition inconsistent when some data are missing. This, and motivation why more advanced weighting methods are useful, is done by presenting Robins' framework in a clear and accessible way, deferring more complex technical details to the Appendix. The Appendix includes the intricate method of double robustness, on which an accessible and insightful perspective is offered. The position of likelihood-based score equations is described as well, pointing to the advantages of efficiency as well as to the dangers that arise from an increased misspecification risk. It would have been nice to see a likelihood-based analysis added to the one presented by the authors, in particular an analysis of the relationship between starting from influence functions, on the one hand, and a full probabilistic specification for likelihood, on the other hand.

2. THE PRETEST–POSTTEST STUDY

The applicability of the results is wider than might be understood from a narrow interpretation of pretest–posttest designs. In fact, most longitudinal settings are embraced by calling the last measurement of interest the posttest measurement and considering intermittent ones as auxiliary measurements, in line with the development in the paper. The work is also important to shed light on the longstanding and still confusing discussion on how to deal with baseline measurements: ignore them, treat them as covariates, treat them as outcomes, subtract them from the measurement of interest and hence provide absolute differences or use them to calculate relative differences instead. The strong advantage of the authors' developments is that their opinion is rooted, not in subjective judgment, school of thought, preference or custom, but rather in the objective results derived from optimality theory. It hopefully will slow down the stream of publications that tackle this issue in an ad hoc, situation-based and subjective fashion. Indeed, one is sometimes under the impression that each team that designs a clinical study feels obliged to reinvent the wheel regarding this issue.

Of course, the developments do not have to be restricted to the pretest–posttest study, not even in its broadest interpretation, since Robins' theory holds very generally. However, by making this deliberate choice, a framework has been selected, which is wide but helps to fix ideas and focus. This contribution will help bring proper incomplete data methodology closer to the board room, even though experience dictates that it still may take a while before it is routinely embraced by, for example, the biopharmaceutical industry and the regulatory authorities, and implemented in clinical trials. However, a remark is necessary here. The drug development process is costly in economic, time consumption and ethical terms. It is therefore imperative to look for the most optimal strategy in *every* aspect of drug development. How then, could one advocate the use of grossly simplistic and incorrect statistical methodology, in conjunction the most sophisticated and advanced biological, molecular, pharmacokinetic, pharmacological and clinical knowledge?

Thus, more than ever, it is necessary to incessantly reiterate that simplistic methods such as LOCF and CC are to be avoided (Mallinckrodt, Clark, Carroll and Molenberghs, 2003; Molenberghs et al., 2004), especially because less than adequate reports of the reverse still abound in certain areas of the scientific literature (Shao and Zhong, 2003). The usefulness of CC, for example, is put to rest in Section 3.2 of the subject article. Of course, a properly weighted version of CC is consistent, but still inefficient. This would seem to be sufficient reason to forget about it altogether, and shift to consistent and optimal or, pragmatically, sufficiently efficient methods.

3. REBUILDING OUR INTUITION

The authors establish very clearly that, under MAR, the use of intermediate measurements, grouped into X_2, is important both for validating the MAR assumptions and for increasing efficiency. While this is intuitively obvious to those familiar with missing data work, it generally is not true, due to the fact that results for situations with complete data are different. This is one of those instances where results under complete (balanced) data differ from their incomplete data counterparts. For example, the basic result for a multivariate normal sample that the mean and variance estimators have independent sampling distributions does not hold under incomplete data, except under missing completely at random (MCAR). It illustrates that MAR and MCAR results do differ in important ways, even

though one would be inclined, especially in a likelihood context, to consider the split between MAR and MNAR (missing not at random) as the really important one. The dependence between mean and variance estimators is related to the fact that $E(Y_2|X_1, Y_1, Z = c) - \mu_2^{(c)}$ provides an intuitively appealing correction to the estimating equations and, hence, in particular to the likelihood equations. It can be seen as an "expected residual" in terms of information obtained on a subject prior to the final measurement Y_2. As a result, it is clear that expected means and observed means need to be different at the posttest occasion when not all measurements are obtained.

4. LIKELIHOOD ANALYSIS OF CASE STUDY

Maximum likelihood is given its proper place as a standard approach under MAR, which nevertheless suffers from the risk of misspecification, since more model elements need to be specified. However, in many settings it is sufficient to specify the joint distribution of $(Y_1, Y_2|X_1, X_2, Z)$ or of $(Y_1, X_2, Y_2|X_1, Z)$, rather than of $V = (X_1, Y_1, X_2, Y_2, Z)$ in full. Such a specification may not be unreasonably difficult in practice, and mild departures from MAR are likely not to distort the inferences terribly much. While this may sound a bit pragmatic, the same is true when it comes to practical implementation of the methods laid out in the article; see, for example, Section 5 on practical implementation, where a pragmatic view is offered, in line with standard modeler practice. Thus, arguably the likelihood approach could be added to the tool kit for the analysis of pretest–posttest designs, together with the methodology advocated by the authors. In fact, when considered jointly, a reasonable route to sensitivity analysis unfolds. The choice between methods is driven by a judgment between sensitivity and efficiency, which can vary from problem to problem. Such a pragmatic attitude is more fruitful than sticking, at all cost and based on principle, to a single mode of analysis. This debate is reminiscent of the Bayesian versus frequentist argument, where a comparable shift from a dogmatic to a pragmatic standpoint has been observed. It is therefore nice that the theory laid out here encompasses both fully parametric and semiparametric models, as detailed in Section 2.2.

To supplement the analyses in Table 1 of the manuscript, our Table 1 presents the results of four likelihood-based longitudinal analyses. For comparison's sake, a completers-only analysis complements the analyses by including all profiles, complete and incomplete. The CD4 profile, made up of the three available CD4 measurements (baseline, 20±5 weeks, 96±5 weeks), is modeled in terms of treatment, CD8 and the same baseline covariates as in Section 1.1. The effect of the baseline covariates is allowed to differ with measurement occasion. Practically, a trivariate normal model—a special case of a linear mixed-effects model (Verbeke and Molenberghs, 2000) with unstructured variance–covariance matrix—is assumed.

TABLE 1
Treatment effect estimates for 96±5 *week CD*4 *counts for ACTG* 175*, based on a longitudinal analysis of the CD*4 *profiles*

Data	Baseline CD4	Estimate	SE
CC	Unadjusted	50.75	11.08
CC	Adjusted	61.59	15.39
All profiles	Unadjusted	59.18	9.98
All profiles	Adjusted	61.14	8.71

NOTE. The analysis includes either the completers only or all available subjects. The treatment effect is based on the difference at 96±5 weeks, adjusted either for covariates or for covariates with, in addition, the baseline CD4 measurement.

Based on this model, the treatment effect at week 96±5 can be considered directly (termed Unadjusted in Table 1, i.e., only covariate adjusted) or after further conditioning the final CD4 measurement on $CD4_0$, which is very easy using standard multivariate normal results. This analysis is termed Adjusted in Table 1, and the corresponding standard error is obtained from the delta method. Note that the estimate of the adjusted analysis, using all profiles, is relatively small compared to all other analyses, which is to be expected. The standard errors from the complete case analyses are high, a concern that augments concern about the method's inconsistency.

5. CONCLUDING REMARKS

Thus, to conclude, the paper convincingly restates that we should forget about the CC t test and other popular methods, shows that IWCC is consistent but inefficient and that the newly proposed method performs best. Using the expectation of Y_2 given other information, in the proper way indicated in the paper, increases efficiency. In this regard, double robustness does not need to be seen as something magic or extravagant, but rather as a carefully picked set of estimating equations that leads to increased efficiency and a decreased risk for an inconsistent result. The authors rightfully refer to this as "estimators for practical use with good properties." In addition, a full likelihood

analysis is possible, but one has to be aware of the misspecification risk. In fact, all "good" methods are based on using information from baseline measurements and covariates, from intermediate covariates as well as intermediate measurement occasions. The (regression) models that capture such information, either to formulate weights, or in a longitudinal or multivariate model in a likelihood context, need to be specified sufficiently correctly.

Through their methodological developments, but in particular also through the illustration with ACTG 175, the authors have shown that correct methods, while not completely trivial, are more than feasible in practice.

Comment

Joseph L. Schafer and Joseph D. Y. Kang

We would like to thank the authors for a well written and thoughtful article. They have given us a clear explanation of the theory of Robins, Rotnitzky and Zhao, especially with regard to considerations of efficiency and double robustness. We also appreciate their willingness to share their data, which has allowed us to evaluate the performance of their new method and compare it to some parametric alternatives.

The crux of the problem is as follows. The parameter of interest, β, is an aspect of the conditional distribution of Y_2 given Z and Y_1, but missingness for Y_2 is related not only to (Z, Y_1) but also to $X = (X_1, X_2)$. Simple procedures like the t test and ANCOVA may be biased because they fail to account for the dependence of R on X. Even if the X variables were not related to R, we would still want to make use of them to improve efficiency because of their ability to predict the missing values of Y_2. Variables that are not really of interest except for the fact that they are potentially correlated with missingness and/or missing outcomes have sometimes been called auxiliary variables (Collins, Schafer and Kam, 2001; Allison, 2002).

How can we use the auxiliary variables? One way is simply to condition on them in the analysis. Under MAR, good estimates of the distribution of Y_2 given Y_1, Z and X are available from the complete cases. For a randomized study, however, conditioning on the postrandomization outcomes X_2 hinders us from making causal inferences about the effect of the treatment. Simple weighting methods such as IWCC model the relationships between the auxiliary variables and the response propensities. Parametric approaches based on maximizing the incomplete-data log-likelihood (Little and Rubin, 2002) or multiple imputation (Rubin, 1987) explicitly model the relationships between X and Y_2. The new method presented in this article models both sets of relationships, but then allows one of these two models to be wrong through the interesting feature of double robustness. The new method is similar in flavor to classical model-assisted procedures for sample surveys, such as ratio and regression estimation; those procedures are most efficient when the underlying model is approximately true but retain their unbiasedness regardless.

While examining these data, we found that the apparent bias in the t-test and ANCOVA estimators is due largely to one variable: off-treatment status. Subjects who went off treatment were six times more likely on the odds scale to have missing values for Y_2 than those who did not. In the control group, the average CD4 count at 96 ± 5 weeks for those who went off treatment was 64 points lower than for those who remained on treatment. In the treament group, the corresponding difference was 110. We are not entirely sure what the off-treatment status variable means, but it appears to measure the subjects' compliance with the assigned regimen. Thus we need to emphasize that the parameter β measures the causal effect of intention to treat (ITT), not the effect of the treatment actually received. This study provides an excellent example of how dropout is often strongly related to noncompliance and how neglecting to account for that relationship can bias the usual ITT estimators (Frangakis and Rubin, 1999). At the same time, it suggests that some alternatives to the ITT effect are worth investigating.

The authors' new method is a big improvement over IWCC, which can never perform very well because it

Joseph L. Schafer is Associate Professor and Joseph D. Y. Kang is Research Assistant, Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, Pennsylvania 16802, USA (e-mail: jls@stat.psu.edu, dyk109@psu.edu).

makes use of X only through its association with R, ignoring the direct relationships between X and Y_2. In fact, the inefficiency of IWCC worsens as the potential for bias grows. As the correlation between X and R becomes stronger, IWCC assigns greater weight to the shrinking pool of respondents whose X values most closely resemble those of the nonrespondents. The IWCC can remove nonresponse bias related to X, but in terms of efficiency it tends to break down when bias corrections are needed most.

One issue that we will address is how the new method compares to parametric alternatives. A parametric approach does not need to model the full joint distribution of (X_1, Y_1, X_2, Y_2, Z). With multiple imputation (MI), we only need to build a reasonable model for Y_2 given X_1, Y_1, X_2 and Z. The computations for MI are simple. Suppose, for example, that we are willing to assume for imputation purposes that $Y_{2i} \sim N(U_i^T \eta, \xi)$, where U_i is a vector of covariates derived from X_{1i}, Y_{1i}, X_{2i} and Z_i. First, we would compute the least-squares estimate $\hat{\eta}$ based on the complete cases. Next, we would draw

$$\xi^* = \left(\sum_i R_i (Y_{2i} - U_i^T \hat{\eta})^2 \right) \Big/ V,$$

where V denotes a random χ^2 variate with degrees of freedom $\sum_i R_i - \dim(U_i)$; next, draw η^* from a normal distribution centered at $\hat{\eta}$ with covariance

$$\xi^* \left(\sum_i R_i U_i U_i^T \right)^{-1}.$$

Finally, draw $Y_{2i}^* \sim N(U_i^T \eta^*, \xi^*)$ for all cases that have $R_i = 0$. Repeating the procedure a small number of times (e.g., ten) produces answers that are reasonably efficient. In practice, the regressors in U_i need to be chosen thoughtfully based on analysis of the complete cases. The assumption of normality is less crucial, because post-imputation analyses like the t test or ANCOVA are not highly sensitive to distributional shape and because this assumption affects only the imputed values of Y_2 rather than the entire sample. Many alternatives to normality are available, such as applying a transformation to Y_{2i}, bootstrap resampling of the empirical residuals $(Y_{2i} - U_i^T \eta^*)$ in the final step of the imputation or switching to a generalized linear model.

Using the authors' data, we imputed the missing values of Y_2 under a normal regression model even though this variable is clearly nonnormal. We used the same mean structure that the authors assumed for $E(Y_2|X_1, Y_1, X_2, Z)$. That is, we fit separate linear models to the treatment and control groups with effects for weight, HIV symptoms, prior antiretroviral therapy, Karnofsky score, off-treatment status, and linear and quadratic effects for CD4 and CD8 at baseline and 20 ± 5 weeks. We generated ten imputations, computed the t-test and ANCOVA estimators for each imputed data set, and combined the results by Rubin's (1987) well-known method for scalar estimands. Estimates from the t test and ANCOVA were 57.86 and 57.22, respectively, with standard errors of 9.48 and 9.62—nearly identical to those from the new method.

We also ran simulations to see how the methods perform over repeated sampling from an artificial population that mimics the observed data but does not correspond exactly to our imputation model or the models used by the authors to compute the $\hat{\pi}$'s and $\hat{e}_q$'s. We created samples in the following way. First, we bootstrapped (X_1, Y_1), drawing these variables from their joint empirical distribution. Next, we set $Z = 1$ with probability 0.75 for each subject, and $Z = 0$ otherwise. Then we generated (X_2, Y_2, R) given (X_1, Y_1, Z) from a sequence of regressions with coefficients chosen to closely resemble estimates from the original sample. Details of these regressions are shown in our Table 1. Note in particular the large effects of off-treatment status on Y_2 and R. By repeated simulation of very large samples ($n = 10^6$), we found that the actual treatment effect in this population is $\beta \approx 53.7$.

For the simulation, we drew 1000 samples of size $n = 2139$, and computed estimates and standard errors by the paired t test, ANCOVA, IWCC, the new method and MI. We were not exactly sure how the authors computed the standard error for IWCC; we could not reproduce their value from the original data, so we decided to omit it. For MI, we imputed the missing values ten times and analyzed the imputed data sets by the paired t test and ANCOVA. The results from these two methods were nearly identical, so we report only those from ANCOVA.

Results from this simulation are summarized in our Table 2. Not surprisingly, the simple paired t-test and ANCOVA estimators are substantially biased. The IWCC, the new method and the MI have no discernible bias even though the underlying models are slightly misspecified. The new method has greater efficiency [lower root mean squared error (RMSE)] than IWCC. The MI estimator is slightly less efficient than the new method, with about 1.5% greater RMSE, but its efficiency can be improved by increasing the number of imputations. Nominal 95% confidence intervals, computed as the estimate plus or minus 1.96 standard

TABLE 1
Population model used in simulations

X_1, Y_1:
(wt, HIV, Karn, prior, CD4_0, CD8_0) $\sim$ empirical distribution

$Z|X_1, Y_1$:
$Z \sim$ Bernoulli(0.75)

$X_2, Y_2, R|X_1, Y_1, Z = 0$:
$\sqrt{\text{CD4}_{20}} \sim$ Normal with mean $6.3 - 1.2$ prior $+ 0.80\sqrt{\text{CD4}_0} - 0.23\sqrt[3]{\text{CD8}_0}$, variance $= 7.3$
$\sqrt[3]{\text{CD8}_{20}} \sim$ Normal with mean $1.59 + 0.19$ prior $- 0.18\sqrt{\text{CD4}_0} + 0.81\sqrt[3]{\text{CD8}_0} + 0.186\sqrt{\text{CD4}_{20}}$, variance $= 0.63$
$\log(P(\text{offtrt} = 1)/P(\text{offtrt} = 0)) = 6.9 - 0.56\text{prior} - 0.043\text{Karn} - 0.07\sqrt[3]{\text{CD8}_0} + 0.09\sqrt[3]{\text{CD8}_{20}} - 0.03\sqrt{\text{CD4}_0} - 0.136\sqrt{\text{CD4}_{20}}$
$\sqrt{\text{CD4}_{96}} \sim$ Normal with mean $0.43\sqrt{\text{CD4}_0} - 0.30\sqrt[3]{\text{CD8}_0} + 0.75\sqrt{\text{CD4}_{20}} - 0.30\sqrt[3]{\text{CD8}_0} - 1.3\text{offtrt}$, variance $= 16.6$
$\log(P(R = 1)/P(R = 0)) = 1.4 - 2.2\text{offtrt}$

$X_2, Y_2, R|X_1, Y_1, Z = 1$:
$\sqrt{\text{CD4}_{20}} \sim$ Normal with mean $6.2 - 0.48\text{HIV} - 1.1\text{prior} + 0.034\text{Karn} + 0.65\sqrt{\text{CD4}_0} - 0.16\sqrt[3]{\text{CD8}_0}$, variance $= 8.9$
$\sqrt[3]{\text{CD8}_{20}} \sim$ Normal with mean $1.79 + 0.003\text{wt} + 0.14\text{HIV} + 0.15\text{prior} - 0.15\sqrt{\text{CD4}_0} + 0.75\sqrt[3]{\text{CD8}_0} + 0.15\sqrt{\text{CD4}_{20}}$, variance $= 0.66$
$\log(P(\text{offtrt} = 1)/P(\text{offtrt} = 0)) = 4.6 - 0.33\text{prior} - 0.027\text{Karn} - 0.19\sqrt[3]{\text{CD8}_0} + 0.19\sqrt[3]{\text{CD8}_{20}} + 0.01\sqrt{\text{CD4}_0} - 0.136\sqrt{\text{CD4}_{20}}$
$\sqrt{\text{CD4}_{96}} \sim$ Normal with mean $-6.4 + 0.02\text{wt} + 0.06\text{Karn} + 0.25\sqrt{\text{CD4}_0} + 0.25\sqrt[3]{\text{CD8}_0} + 0.82\sqrt{\text{CD4}_{20}} - 0.60\sqrt[3]{\text{CD8}_0}$
$- 2.5\text{offtrt}$, variance $= 12.7$
$\log(P(\text{R} = 1)/P(\text{R} = 0)) = -1.0 + 0.4\text{HIV} + 0.03\text{Karn} + 0.08\sqrt[3]{\text{CD8}_0} - 0.10\sqrt[3]{\text{CD8}_{20}} - 0.05\sqrt{\text{CD4}_0} + 0.02\sqrt{\text{CD4}_{20}} - 1.9\text{offtrt}$

errors, have actual coverage close to 95% for both the new method and MI. On average, the MI intervals are a bit wider than the new method's, due again to the fact that we are using only ten imputations.

The next question we considered is how the new method and MI respond to greater degrees of model misspecification. For the new method we removed the important off-treatment (offtrt) status variable first from the computation of the $\hat{\pi}$'s, then from the $\hat{e}_q$'s, then from both. For MI we removed this variable from the imputation model. The performance of the modified procedures is summarized in our Table 3. When offtrt is removed from the $\hat{\pi}$ model or the $\hat{e}_q$ model alone, the new method still performs quite well, showing that the double robustness property works as it should. When offtrt is removed from both, the new method is biased, and MI without offtrt is biased to about the same degree. These biases are enough to drop the simulated coverage of the intervals to about 90%, which corresponds to a doubling of the Type 1 error rate in a 0.05-level test.

TABLE 2
Results from samples of size $n = 2139$ for average estimate (true $\beta \approx 53.7$), root mean squared error, percent coverage of nominal 95% interval and average interval width

	t test	ANCOVA	IWCC	New	MI
Avg	62.9	61.5	53.7	53.8	53.9
RMSE	13.9	12.9	11.5	10.2	10.4
Coverage	83.4	87.3	—	93.0	95.8
Width	79.1	80.3	—	75.2	84.8

In some respects, this example is different from those we have typically seen in the social and behavioral sciences, because the treatment effect is large and highly significant regardless of what we do. The signal-to-noise ratio is so large that the biases in the naive methods that do not use the auxiliary variables (*t* test and ANCOVA) amount to a standard error or more. In our experience it is a bit unusual to find auxiliary variables that are correlated with the missingness indicator R to the degree exhibited in these data by the variable offtrt. Even if such variables are present, we have found that they usually do not interact with the covariates of interest (in this case, Z and Y_1) and the response (Y_2) strongly enough to seriously degrade the performance of intervals and tests, except in situations with unusually large n and very high power. In situations with less power, the real advantage in using auxiliary variables is not to reduce bias, but to increase efficiency. For ex-

TABLE 3
Results from samples of size $n = 2139$ with the off-treatment status variable removed from missing-data procedures

	New*	New†	New*†	MI†
Avg	53.9	53.8	58.6	59.2
RMSE	10.2	10.2	10.8	11.3
Coverage	91.9	93.9	89.9	90.4
Width	71.4	76.7	74.0	78.0

NOTE. New* removes it from computation of $\hat{\pi}$'s; New† removes it from computation of $\hat{e}_q$'s; New*† removes it from computation of both $\hat{\pi}$'s and $\hat{e}_q$'s; and MI† removes it from the imputation model.

ample, when subjects drop out of a longitudinal study, intermediate measurements can be quite valuable for predicting missing endpoints even if the dropout is completely at random.

To see how the methods perform in a situation with more noise, we repeated our simulation with a reduced sample size of $n = 400$; this leaves us with an average of 100 subjects in the control group and 300 in the combined treatment groups at baseline. Under these conditions the power of an ordinary 0.05-level paired t test with no dropout is about 85%, which seems plausible for a randomized trial. The results from this new simulation are summarized in our Table 4. The paired-t and ANCOVA estimators are just as biased as they were with $n = 2139$ in absolute terms, but these biases are now less consequential because the standard errors are larger. The new method still has essentially no bias, but it is less efficient than the "naive" methods and its coverage has begun to suffer. In this situation the expected number of respondents in the control and treatment groups is about 60 and 185, respectively. One might think that samples of this size are large enough for a robust comparison of two means, but the asymptotic approximations of the new method seem to require samples larger than those to which we are ordinarily accustomed. With samples of $n = 400$, the MI method still works well. This is consistent with what we have found in other simulations—although technically a large-sample procedure, it can work very well with moderate or small n (e.g., Graham and Schafer, 1999).

TABLE 4
Results from samples of size $n = 400$

	t test	ANCOVA	IWCC	New	MI
Avg	61.6	60.2	52.5	52.5	52.6
RMSE	24.9	24.5	29.8	26.8	25.4
Coverage	92.3	93.8	—	90.7	96.8
Width	183	186	—	180	218

In summary, this new method requires a plausible model for either the response propensities or the $X - Y_2$ relationships, and it will be most efficient when the latter is approximately true. Multiple imputation needs a plausible model for the $X - Y_2$ relationships to have low bias and high efficiency. Either way, modeling of $E(Y_2|X_1, Y_1, X_2, R)$ is a good idea and should not be done haphazardly. The new method requires us to fit three sets of regressions, whereas MI requires one set of regressions plus simulation of random variates. The new method also requires the samples to be rather large. Under the right conditions the new method performs beautifully and we can wholeheartedly recommend it.

ACKNOWLEDGMENT

This research was supported by Grant 1-P50-DA10075 from the National Institute on Drug Abuse.

Rejoinder

Marie Davidian, Anastasios A. Tsiatis and Selene Leon

INTRODUCTION

We thank all of the discussants for their thoughtful and insightful comments. We very much enjoyed reading all of the discussions and we have learned a great deal more about the interrelationships among different perspectives on missing data problems from them. While all discussions touch on some common themes, each one also raises some different, important issues. Accordingly, we respond to the discussants' comments in turn, focusing mostly on several of these issues. The relative length and extent of our responses to each discussion by no means reflect the relative importance of the comments.

RESPONSE TO AN AND LITTLE

An and Little take the position that one may appeal to existing principles of regression modeling (PSM) as the basis for methods for pretest–posttest analysis, with or without data MAR. We do not disagree that methods based on regression modeling are a useful approach to these and more general problems. However, we believe that the distinction between the methods that emerge from application of the semiparametric theory of Robins, Rotnitzky and Zhao (1994, RRZ) and PSM methods is less profound than the debates in the literature, which tend to feature "schools" (in the words of Molenberghs) that advocate one approach or another and imply a sort of mutual exclusivity of the

methods, would suggest. Our perspective is that appealing to the semiparametric theory can in fact highlight and clarify formally the interrelationships among methods. Indeed, the semiparametric theory characterizes the class of all consistent (RAL) estimators for β, including the efficient one within this class. Accordingly, we expect that many standard "nice" estimators may be represented as members of this class. Thus, we believe it is fruitful and instructive to view all methods from the perspective of this theory, a theme we highlight throughout.

An and Little have very nicely summarized in the case where there are no intervening covariates X_2 what can be expected via a PSM approach both with and without missing data. As they point out, under randomization methods for estimation of $\beta = E(Y_2|Z = 1) - E(Y_2|Z = 0)$ based on least squares (LS) fitting of a (parametric) regression model for $E(Y_2|X_1, Y_1, Z)$ will yield consistent inference even if the model is incorrect. Although this can be deduced from considering the regression model directly, it also can be seen to follow from the semiparametric theory of influence functions. For example, if we assume

$$(1)\quad E(Y_2|X_1, Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 X_1 + \beta Z,$$

the usual ANCOVA approach supplemented by adjustment for additional baseline covariates, the LS estimator for β in (1) may be shown to have influence function in the class given in (3) of our paper and, hence, is consistent for $\beta = E(Y_2|Z = 1) - E(Y_2|Z = 0)$ even if (1) does not correspond to the true regression relationship.

When Y_2 is MAR and there are no intervening covariates X_2, An and Little remind us that the incomplete cases do not contain information on the regression $E(Y_2|X_1, Y_1, Z)$, so if our interest is on the regression relationship only, we may base inference on the complete cases. However, the quantity of interest, $\beta = E(Y_2|Z = 1) - E(Y_2|Z = 0)$, which may be derived from this regression, also depends on the distribution of (X_1, Y_1, Z), which must be deduced from all the data, as also noted by Brumback and Brumback. Hence, a potential pitfall is that an incorrect regression model for the complete cases can lead to bias, a point to which An and Little allude and one we feel is worth demonstrating explicitly. For simplicity, assume no additional covariates X_1 and that Y_1 is binary. Suppose we postulate

$$(2)\qquad E(Y_2|Y_1, Z) = \alpha_0^* + \alpha_1^* Y_1 + \beta Z$$

and propose to estimate β via the LS estimator for β in (2). Suppose in truth

$$(3)\qquad E(Y_2|Y_1, Z) = \alpha_0 + \alpha_1 Y_1 + \gamma Z + \kappa Z Y_1,$$

which implies that $\beta = E(Y_2|Z = 1) - E(Y_2|Z = 0) = \gamma + \kappa\rho$, where $\rho = P(Y_1 = 1|Z) = P(Y_1 = 1)$ by randomization. Suppose further that $P(R = 1|Y_1, Y_2, Z) = P(R = 1|Y_1)$ (the MAR assumption, with missingness dependent on Y_1 nondifferentially by treatment group) and that $P(R = 1|Y_1 = y) = \pi_y$, $y = 0, 1$. Then it is straightforward to show that the LS estimator for β in (2) based on the complete cases only converges in probability under the true relationship (3) not to the quantity of interest, $\beta = \gamma + \kappa\rho$, but to

$$\gamma + \kappa\left\{\frac{\rho\pi_1}{\pi_0(1-\rho) + \pi_1\rho}\right\}.$$

That is, the LS estimator based on the incorrect model (2) estimates a quantity that differs from that of interest by an amount that has to do with the difference between the population proportion $\rho = P(Y_1 = 1)$ and $(\rho\pi_1)/\{\pi_0(1-\rho)+\pi_1\rho\} = P(Y_1 = 1|R = 1)$, the proportion among complete cases. Thus, this estimator fails to incorporate required information on Y_1 from the entire population, leading to inconsistency. This example highlights that a potential price of a pure PSM approach is inconsistent inference. From the view of the semiparametric theory, as such an estimator is inconsistent, it is not a member of the class of all (consistent) RAL estimators; hence it would not emerge as a candidate for inference on β.

As An and Little discuss clearly, when missingness depends on intervening covariates, it is necessary to incorporate the information on them from the incomplete cases, but it is not appropriate to simply include these covariates in a regression model for the outcome. They note that the general parametric PSM approach may be implemented by obtaining imputed/predicted missing responses $\widehat{e}_{q(1)i}$ based on a model for $E(Y_2|X_1, Y_1, X_2, Z) = E(Y_2|X_1, Y_1, X_2, Z, R = 1)$ (so using complete cases only) and carrying out a regression analysis to estimate β under a model for $E(Y_2|X_1, Y_1, Z)$ (which may in fact be incorrect as above), substituting the imputed values for the missing responses. For example, for the single mean $\mu_2^{(1)}$ and a linear model for $E(Y_2|X_1, Y_1, X_2, Z = 1)$, the estimator is

$$(4)\quad \widehat{\mu}_2^{(1)} = n^{-1}\left\{\sum_{i=1}^{n} R_i Y_{2i} + \sum_{i=1}^{n}(1 - R_i)\widehat{e}_{q(1)i}\right\}.$$

Regression imputation or multiple imputation (discussed by Schafer and Kang) may be used to obtain the $\widehat{e}_{q(1)i}$. They note that this requires $E(Y_2|X_1, Y_1, X_2, Z)$ be correctly specified, raising concern over the potential for bias associated with an incorrect parametric model. As An and Little point out, nonparametric regression may address this concern at the expense of being subject to the "curse of dimensionality," leading them to propose the propensity spline prediction (PSP) method (Little and An, 2004), in which the imputation is based on, roughly, a model of the form, in the case of $\mu_2^{(1)}$, which we focus on here, $E\{Y_2|\pi(X_1, Y_1, X_2, Z = 1), A\}$, where A is a function of (X_1, Y_1, X_2) and this is estimated using splines. Brumback and Brumback also discuss regression on the propensity π as a means of reducing dimensionality.

To illuminate the connection between PSP and the class of RAL estimators derived in our paper, we reiterate that our proposed estimators follow from making no assumptions on the joint distribution of (X_1, Y_1, X_2, Y_2, Z) beyond independence of (X_1, Y_1) and Z along with an assumption on the form of, for example, in the case of $\mu_2^{(1)}$, $\pi^{(1)}(X_1, Y_1, X_2)$. Although in this case correct modeling of $E(Y_2|X_1, Y_1, Z = 1)$ and $E(Y_2|X_1, Y_1, X_2, Z = 1)$ serves to enhance efficiency, no assumptions on these regression relationships are required and, as long as the assumption on $\pi^{(1)}$ is correct, consistency is obtained regardless of whether the regressions are modeled correctly. In the PSP approach, An and Little make an assumption on $\pi^{(1)}(X_1, Y_1, X_2)$ but make no assumptions on regression relationships such as $E\{Y_2|\pi(X_1, Y_1, X_2, Z = 1), A\}$, instead modeling these nonparametrically. Thus, PSP is derived under the same conditions as the estimators in our paper. Accordingly, if PSP estimators are RAL, they must have influence functions in the class of influence functions given by the RRZ theory and hence must be in the resulting class of estimators.

In a simple special case where there is only one variable, we can show easily that the PSP estimator has influence in the class that corresponds to consistent RAL estimators. Suppose we consider just the data for one treatment group, have only (Y_1, Y_2), and focus on estimation of $E(Y_2) = \mu_2$. Assume $P(R = 1|Y_1, Y_2) = P(R = 1|Y_1) = \pi(Y_1)$ is known and discrete, taking on values $\pi_1, \ldots, \pi_K$, say. Under these conditions the natural nonparametric estimator for $E\{Y_{2i}|\pi(Y_{1i}) = \pi_j\}$ is

$$\widehat{E}_j = \frac{\sum_{i:\pi(Y_{1i})=\pi_j} R_i Y_{2i}}{\sum_{i:\pi(Y_{1i})=\pi_j} R_i},$$

yielding the estimator for μ_2 found by substituting in (4) $\widehat{e}_{q(1)i} = \widehat{E}_j$ if $\pi(Y_{1i}) = \pi_j$, which reduces to $\widehat{\mu}_2 = n^{-1}\sum_{j=1}^{K} r_j \widehat{E}_j$, $r_j = \sum_{i=1}^{n} I\{\pi(Y_{1i}) = \pi_j\}$. This may be rewritten as

$$\begin{aligned} &n^{-1}\left[\sum_{j=1}^{K}\sum_{i:\pi(Y_{1i})=\pi_j}\left\{\frac{R_i(Y_{2i} - \widehat{E}_j)}{\pi_j} + \widehat{E}_j\right\}\right] \\ &= n^{-1}\sum_{i=1}^{n}\left[\frac{R_i\{Y_{2i} - \widehat{e}_{q(1)i}\}}{\pi(Y_{1i})} + \widehat{e}_{q(1)i}\right] \qquad (5) \\ &= n^{-1}\sum_{i=1}^{n}\left[\frac{R_i Y_{2i}}{\pi(Y_{1i})} - \frac{R_i - \pi(Y_{1i})}{\pi(Y_{1i})}\widehat{e}_{q(1)i}\right]. \end{aligned}$$

The first term in the second expression in (5) is the "calibration" correction to which An and Little refer (which in fact equals 0 in this simple example, but need not in general). This estimator has influence function

$$\frac{R(Y_2 - \mu_2)}{\pi(Y_1)} - \frac{R - \pi(Y_1)}{\pi(Y_1)}E\{Y_2|\pi(Y_1)\},$$

which is of the form of those following from the RRZ semiparametric theory for estimators for a single mean. Thus, we may conclude immediately that $\widehat{\mu}_2$ is consistent for μ_2 and asymptotically normal with asymptotic variance that may be deduced from the influence function.

It is important to recognize that this influence function and those corresponding to PSP estimators in more general settings belong to the class of influence functions for semiparametric RAL estimators because a nonparametrically consistent estimator for, in this case, $E\{Y_2|\pi(Y_1)\}$ is used. If a parametric model were used here, this would impose additional assumptions beyond those of a semiparametric model. By working directly with the class of influence functions indicated by the theory, for example, in the simple example

$$\frac{R(Y_2 - \mu_2)}{\pi(Y_1)} - \frac{R - \pi(Y_1)}{\pi(Y_1)}g(Y_1),$$

one has greater latitude to choose $g(Y_1)$ to develop consistent estimators.

In the general setup of our paper, whether the semiparametric efficient estimator may be represented as a PSP estimator or, equivalently, whether a PSP estimator may be shown to have the efficient influence function, is not readily clear and would be interesting to establish. An and Little contend that PSP addresses the curse of dimensionality, but this is only true when the propensity π is correctly specified; indeed, to specify π correctly also involves a curse. When $\pi^{(1)}$ is correctly modeled, basing inference instead on the class

of estimators in our paper ensures that the analyst does not have to worry about the curse in the sense that, while the construction of the PSP estimators requires that regression relationships be modeled nonparametrically for consistency, the form and double robustness property of our estimators ensure consistency even if regression relationships are represented by incorrect models.

Viewing the PSP approach as a way to implement estimators in the class following from the RRZ theory, then we wonder whether it may in fact have some pleasing empirical properties. Estimators in the class, whose form incorporates inverse weighting explicitly, can be numerically unstable when some cases have associated small values of $\pi^{(1)}$. The PSP representation may lessen this effect, as suggested by the comment of An and Little that PSP "obviates the need for" the calibration correction.

Overall, however, it is notable that our estimators have a simple closed form that requires only that the analyst carry out familiar modeling exercises. Moreover, because the influence functions of our estimators are readily available, calculation of closed form standard errors via the sandwich method is immediate.

An and Little express surprise at our statements that there is no general consensus on appropriate approaches to pretest–posttest analysis, particularly in the face of missing data, and that some of the approaches they discuss are often not used by practitioners. We agree that many practitioners are well aware of the issues raised by An and Little and are indeed basing their inferences on sound and sophisticated principles. Our experience in the clinical trial, pharmaceutical and regulatory settings, however, more closely mirrors that of Molenberghs. Statistical sections of study protocols that propose complete case (CC) analyses are commonplace in our experience; for example, the protocol for a recent HIV study states that the primary analysis will be based on "change in CD4+ cell count from baseline" to the regular follow-up visit at 32 months (so based on the unfortunately named paired t-test method) and that "... patients who are lost to follow-up will not contribute to this comparison." In our collaborations, we have routinely witnessed debates over whether methods based on change scores (posttest–pretest) or ANCOVA should be used (a point addressed by Brumback and Brumback), whether failure of the response to follow a normal distribution will bias results, whether adjustment for additional baseline covariates should even be undertaken and whether CC or last observation carried forward (LOCF) is the more appropriate approach to handling missing follow-up responses.

Overall, we believe that semiparametric theory can shed considerable insight on this and other problems, and can suggest not only estimators that may be alternatively motivated from a PSM perspective, but also provide a formal framework in which to view these and many other nice estimators. Through the lens of this theory, one can observe that many seemingly disparate approaches share common themes.

RESPONSE TO BRUMBACK AND BRUMBACK

Brumback and Brumback first emphasize an important point that has been the source of some misconception among practitioners, namely, that basing a pretest–posttest analysis on change scores is not the same as basing it on a method that performs a regression adjustment for Y_1. In doing so, they make an interesting connection between this phenomenon, known as Lord's paradox, and the celebrated Simpson paradox, which has to do with difficulties with confounding that arise in, for instance, epidemiological studies, highlighting the link between causal inference and inference under MAR.

Brumback and Brumback bring up an intriguing alternative approach, which they discuss in the context of estimation of $\mu_2^{(1)}$ and the special case of no missing data, no baseline or intervening covariates (X_1, X_2) and binary Y_1. In general, from (3) and (4) of our paper, the SPEE for $\mu_2^{(1)}$ is

$$\widehat{\mu}_2^{(1)} = n^{-1} \sum_{i=1}^{n} \{ Z_i Y_{2i} - (Z_i - \widehat{\delta}) \widehat{e}_{h(1)i} \}, \tag{6}$$

where $\widehat{\delta} = n_1/n$ and $\widehat{e}_{h(1)i}$ is the predicted value for i based on an estimator for $E(Y_{2i}|Y_{1i}, Z_i = 1)$. In the particular case where Y_1 is binary, as Brumback and Brumback point out, the obvious estimators are $\overline{Y}_{20}$ for $E(Y_2|Y_1 = 0, Z = 1)$ and $\overline{Y}_{21}$ for $E(Y_2|Y_1 = 1, Z = 1)$. Brumback and Brumback contend that the SPEE in this setting is given by

$$\widetilde{\mu}_2^{(1)} = n_1^{-1} \sum_{i=1}^{n} Z_i Y_{2i} = m_0 \overline{Y}_{20}/n_1 + m_1 \overline{Y}_{21}/n_1,$$

which weights the estimators for $E(Y_2|Y_1 = y, Z = 1)$, $y = 0, 1$, by within-treatment proportions $w_0 = m_0/n_1$ and $w_1 = m_1/n_1$. However, substituting $\overline{Y}_{20}$ and $\overline{Y}_{21}$ in (6) followed by algebra shows that the SPEE is in fact given by

$$\widehat{\mu}_2^{(1)} = r_0 \overline{Y}_{20}/n + r_1 \overline{Y}_{21}/n, \tag{7}$$

where $r_0 = \sum_{i=1}^{n}(1 - Y_{1i})$ and $r_1 = \sum_{i=1}^{n} Y_{1i}$, which weights by overall proportions. A similar expression obtains for $\mu_2^{(0)}$. We highlight this to emphasize the important point that, even though we focus here on the mean for the single treatment group with $Z = 1$, the SPEE gains efficiency by exploiting the information from both treatment groups.

This issue aside, Brumback and Brumback raise an interesting possibility, that of a so-called compromise estimator (CE) based on shrinkage ideas. The version given in the discussion could likely be modified to exploit information from both treatments. When τ^2 is known, the expression for the CE has smaller variance than (7), which is the sample mean for $Z = 1$ if $\tau^2 > 0$ and has variance equal to that of (7) if $\tau^2 = 0$. As τ^2 would be unlikely to be known in practice but evidently was taken as such in the MSE comparisons presented by the authors [the "estimate" of τ^2 is the variance of the chosen values of $E(Y_2|Y_1 = 0, Z = 1)$ and $E(Y_2|Y_1 = 1, Z = 1)$, which would not be known in practice], we do not have a sense of the extent to which the need to estimate τ^2 would impact practical performance of the CE relative to (7) [or (6)] of this rejoinder. We conjecture that the CE (with τ^2 estimated realistically from the data) may be a superefficient estimator and hence is not regular; accordingly, it is excluded from the class of RAL estimators for $\mu_2^{(1)}$ to which the influence functions in (3) of our paper correspond. Nonetheless, this is not to say that it may not have desirable properties. The CE with τ^2 estimated is an intriguing idea that we believe deserves further study.

Turning to the issue of handling intervening covariates X_2 when Y_2 is MAR, ignoring for simplicity X_1, Y_1 and from a perspective similar to that taken by An and Little, Brumback and Brumback demonstrate that $\mu_2^{(1)} = E(Y_2|Z = 1)$ (and similarly $\mu_2^{(0)}$) may be estimated by averaging $E(Y_2|X_2, Z = 1, R = 1)$ over the distribution of $X_2|Z = 1$, that is, over the entire population, not just among those with Y_2 missing. They then propose a PSM approach based on regression modeling on the propensity score in a spirit similar to that of An and Little, with the additional twist of using the shrinkage-based CE for the regression modeling. However, this approach is not equivalent to the PSP of An and Little, because Brumback and Brumback base inference on their equations (3) and (4) rather than an equation like our (4) in this rejoinder, and we are uncertain as to how they implemented estimation of the regression relationships (i.e., parametric or nonparametric modeling).

Brumback and Brumback end by posing several questions. First, they question why we focus on an intent-to-treat estimand. We do not disagree that an analysis focused on treatment efficacy would be of interest. However, our emphasis on intent-to-treat reflects that this would be the standard analysis in the clinical trial, pharmaceutical and regulatory setting. In the event where noncompliance in fact leads to missingness (e.g., dropout), this view may be interpreted as focusing on the estimand that would be of interest if there were no missing data, and the analysis may then be interpreted as attempting to estimate this quantity in the unfortunate circumstance that dropout did occur.

We are not entirely clear as to the motivation for the second point raised by Brumback and Brumback. In general, if one factorizes a likelihood in terms of $W|V$ and V, the component that corresponds to V is orthogonal to the first term in the sense that parameters are variation independent, and then all resulting influence functions for estimators for a parameter in the first term have influence functions that satisfy the condition $E\{\varphi(W)|V\}$. Under these conditions, we would indeed recommend a conditional analysis.

Finally, Brumback and Brumback ask about our practical recommendation to include covariates in the regression models involved in computation of the proposed estimator. Mathematically, including covariates should increase efficiency, which can be appreciated from a geometric perspective, because the influence function can be viewed as a projection onto a linear space spanned by the covariates. As the size of that space increases, the projection becomes smaller and hence has smaller variance. However, Brumback and Brumback raise the important point that there is a threshold in practical problems above which including additional, potentially unnecessary covariates in the models will lead to instability in smaller sample sizes (a point raised also by Schafer and Kang). An interesting question for future research is the rate at which one should increase model complexity relative to sample size. Brumback and Brumback end by posing several intriguing questions for future research, which we cannot hope to address in this limited space.

RESPONSE TO MOLENBERGHS

We agree wholeheartedly with virtually all of Molenberghs' comments. He has presented with considerably more eloquence than we could hope to achieve our position on handling missing data in practice, in general, and the pretest–posttest problem, in

particular. He emphasizes that the continued, erroneous use of CC and LOCF analyses is likely in part a consequence of the existence of competing "schools" in the literature, a point with which we concur, and he provides compelling arguments to support our position that viewing methods pragmatically from the perspective of semiparametric theory can lead to considerable insight.

Molenberghs explicitly discusses likelihood analysis, which we did not emphasize in our paper, providing yet another complementary perspective. He correctly points out that our implication that likelihood methods require specification of the full joint distribution of (X_1, Y_1, X_2, Y_2, Z) is an overstatement; indeed, only aspects of this distribution must be specified (but must be specified correctly). In fact, one way to contrast our approach based on the RRZ theory to that of maximum likelihood is alluded to by Molenberghs. As we noted in our response to An and Little, the semiparametric RRZ theory takes the point of view that one is willing to make assumptions on the probabilities of observing Y_2 but not on regression relationships for Y_2, an approach that leads to the double robustness property if one characterizes the regression relationships correctly and to consistent inference regardless. In a maximum likelihood approach, one instead makes assumptions on regression relationships such as those noted by Molenberghs, and in fact need not even make any assumptions on the $\pi^{(c)}(X_1, Y_1, X_2)$, $c = 0, 1$. However, this comes at a price, because the regression relationships need to be specified, in the words of Molenberghs, "sufficiently correctly" to achieve unbiased inference.

RESPONSE TO SCHAFER AND KANG

Schafer and Kang provide illuminating and helpful perspectives on several issues. Like Brumback and Brumback, they also raise the point of our focus on the intent-to-treat estimand and provide an excellent discussion of the biases that can arise when noncompliance is related to dropout. Schafer and Kang also make the connection between the semiparametric methods we discuss and methods in the sample survey literature, which are, in fact, based on the same ideas, but which evolved from an entirely different perspective. Schafer and Kang also offer a very useful and intuitive explanation of the suboptimal performance of the IWCC estimator relative to that of the proposed estimators based on the efficient influence function.

A welcome contribution by Schafer and Kang is the extensive set of simulation studies they present based on the ACTG 175 scenario that compares the proposed approach not only to popular estimators directly, but to one version (ANCOVA) where the missing Y_2 were filled in via multiple imputation (MI). In doing so, they note in the same spirit as Molenberghs' remark on maximum likelihood that the latter approach does not require full specification of the joint distribution of (X_1, Y_1, X_2, Y_2, Z), which the remark in our paper erroneously suggested. The simulations illustrate several important points. Under ideal conditions (e.g., correct modeling), the proposed method based on the efficient influence function and the MI method achieve similar performance, with perhaps a slight edge to the proposed method, which echoes Molenberghs' view that under a pragmatic approach all "good" methods should yield similar inferences. Further simulations exhibit convincingly both the double robustness property and the potential for bias of the proposed approach when both $\pi^{(c)}$ and regression relationships for Y_2 are modeled incorrectly, and of the MI approach when the imputation model is incorrectly specified.

Schafer and Kang also report on a simulation that addresses the spirit of the comment by Brumback and Brumback regarding performance in smaller samples. The simulations with $n = 400$ demonstrate a potential pitfall of the proposed methods, namely, that practical performance can be degraded when model complexity is fairly high and sample size is not too large. This prompts us to issue a cautionary note that the operating characteristics of inverse-weighted methods in this setting, not only for the pretest–posttest problem, but when applied in other problems, need to be better understood. We conjecture that this is of particular concern when some of the $\pi^{(c)}$ are very small. As noted above, the PSP approach to implementing estimators in this class advocated by An and Little may offer better practical performance.

CLOSING COMMENTS

We would again like to offer a strong vote of thanks to all the discussants. Their incisive comments have enhanced tremendously the message and utility for practitioners we hope to achieve with this paper and raised many issues for further research.

In closing, we would like to bring up one additional issue that did not arise in any of the discussions. All the methods discussed here rely on the validity of the MAR assumption. In settings where Y_2 is missing exclusively due to dropout, the analyst may feel confident adopting this assumption when sufficient information

on reasons for dropout (e.g., X_1, X_2) is available. One situation in which the MAR assumption would be suspect is in the case where Y_2 is missing due to the intervening death of a subject, where missingness due to death may be related to underlying disease state. More fundamentally, this scenario raises the philosophical issue of what a reasonable question of interest regarding the response really is. In the setting of ACTG 175, for example, where Y_1 and Y_2 are CD4 count, it is natural to ask the meaning of CD4 count if a subject has died. If death is due to HIV, then it is not clear what CD4 count at a subsequent time represents, whereas, in contrast, one may still envision CD4 postdeath for a subject who, for example, died due to accidental causes. In the former case, one might argue that, as diminishing CD4 is strongly associated with poor prognosis with presumably no detectable CD4 count corresponding to complete annihilation of the immune system, taking CD4 as equal to 0 for subjects whose death is clearly related to HIV might be a biologically defensible solution. Nonetheless, this seems somewhat unsatisfactory and may not be applicable in other situations. In ACTG 175, of the 739 subjects missing Y_2, only 49 in fact died prior to 96 weeks. Accordingly, we took the pragmatic view that ignoring this inconvenient feature would not detract too much from asking a question about CD4 at 96 ± 5 weeks for the vast majority of subjects who did not die. In general, the analyst needs to think carefully about the implications of death and come to a satisfactory resolution on interpretation of the effect of interest on a case-by-case basis.

We end by reiterating that we do not wish to suggest that the proposed methods should supplant all others as the methods of choice for pretest–posttest analysis. Our objective for the paper was to demystify the RRZ theory for practitioners and demonstrate how the theory lends insight into the structure of the problem and the interrelationships among approaches. Pragmatic data-analytic techniques applied in any of the approaches should lead to correct and relatively efficient inference, a point we believe the RRZ theory helps to solidify.

ADDITIONAL REFERENCES

ALLISON, P. D. (2002). *Missing Data*. Sage, Thousand Oaks, CA.

CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89** 81–87.

COLLINS, L. M., SCHAFER, J. L. and KAM, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods* **6** 330–351.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11** 89–121.

FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379.

GRAHAM, J. W. and SCHAFER, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In *Statistical Strategies for Small Sample Research* (R. Hoyle, ed.) 1–29. Sage, Thousand Oaks, CA.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

LITTLE, R. J. A. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

LORD, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin* **68** 304–305.

MALLINCKRODT, C. H., CLARK, W. S., CARROLL, R. J. and MOLENBERGHS, G. (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J. Biopharmaceutical Statistics* **13** 179–190.

MOLENBERGHS, G., THIJS, H., JANSEN, I., BEUNCKENS, C., KENWARD, M. G., MALLINCKRODT, C. and CARROLL, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5** 445–464.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27** 85–95.

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

SHAO, J. and ZHONG, B. (2003). Last observation carry-forward and last observation analysis. *Statistics in Medicine* **22** 2429–2441.

STONE, R. (1993). The assumptions on which causal inferences rest. *J. Roy. Statist. Soc. Ser. B* **55** 455–466.

VAN BUUREN, S. and OUDSHOORN, C. G. M. (1999). Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054. For associated software see www.multiple-imputation.com.

VERBEKE, G. and MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.