Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models

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

Consider a study whose design calls for the study subjects to be followed from enrollment (time $t = 0$) to time $t = T$, at which point a primary endpoint of interest Y is to be measured. The design of the study also calls for measurements on a vector $\mathbf{V}(t)$ of covariates to be made at one or more times t during the interval $[0, T)$. We are interested in making inferences about the marginal mean μ_0 of Y when some subjects drop out of the study at random times Q prior to the common fixed end of follow-up time T. The purpose of this article is to show how to make inferences about μ_0 when the continuous drop-out time Q is modeled semiparametrically and no restrictions are placed on the joint distribution of the outcome and other measured variables. In particular, we consider two models for the conditional hazard of drop-out given $(\bar{\mathbf{V}}(T), Y)$, where $\bar{\mathbf{V}}(t)$ denotes the history of the process $\mathbf{V}(t)$ through time $t, t \in [0, T)$. In the first model, we assume that $\lambda_Q(t|\bar{\mathbf{V}}(T), Y) = \lambda_0(t|\bar{\mathbf{V}}(t)) \exp(\alpha_0 Y)$, where α_0 is a scalar parameter and $\lambda_0(t|\bar{\mathbf{V}}(t))$ is an unrestricted positive function of t and the process $\bar{\mathbf{V}}(t)$. When the process $\bar{\mathbf{V}}(t)$ is high dimensional, estimation in this model is not feasible with moderate sample sizes, due to the curse of dimensionality. For such situations, we consider a second model that imposes the additional restriction that $\lambda_0(t|\bar{\mathbf{V}}(t)) = \lambda_0(t) \exp(\boldsymbol{\gamma}_0' \mathbf{W}(t))$, where $\lambda_0(t)$ is an unspecified baseline hazard function, $\mathbf{W}(t) = \mathbf{w}(t, \bar{\mathbf{V}}(t))$, $\mathbf{w}(\cdot, \cdot)$ is a known function that maps $(t, \bar{\mathbf{V}}(t))$ to R^q, and $\boldsymbol{\gamma}_0$ is a $q \times 1$ unknown parameter vector. When $\alpha_0 \neq 0$, then drop-out is nonignorable. On account of identifiability problems, joint estimation of the mean μ_0 of Y and the selection bias parameter α_0 may be difficult or impossible. Therefore, we propose regarding the selection bias parameter α_0 as known, rather than estimating it from the data. We then perform a sensitivity analysis to see how inference about μ_0 changes as we vary α_0 over a plausible range of values. We apply our approach to the analysis of ACTG 175, an AIDS clinical trial.

KEY WORDS: Augmented inverse probability of censoring weighted estimators; Cox proportional hazards model; Identification; Missing data; Noncompliance; Nonparametric methods; Randomized trials; Sensitivity analysis; Time-dependent covariates.

1. INTRODUCTION

Rotnitzky, Robins, and Scharfstein (1998) proposed augmented inverse probability of censoring weighted (AIPCW) semiparametric estimators for the marginal mean μ_0 of an outcome of interest Y measured at a fixed time T from longitudinal data when (a) some subjects drop out of the study, (b) drop-out is nonignorable, and (c) the probability of drop-out is a function of the potentially unobserved Y and additional time-dependent covariates $\mathbf{V}(t)$ and follows a parametric model. The Rotnitzky–Robins–Scharfstein AIPCW estimators are semiparametric in the sense that they are guaranteed to be consistent and asymptotically normal (CAN) for μ_0 regardless of the joint distribution of the outcome Y and the additional variables $\mathbf{V}(t)$, provided that the parametric model for drop-out is correct. A natural extension of their approach is to allow the model for drop-out to be semiparametric. This generalization will be particularly important in studies in which time to drop-out is a continuous random variable. In particular, to ensure additional robustness to model misspecification, we allow the time to drop-out to depend on Y and other possibly time-dependent variables $\mathbf{V}(t)$ through a semiparametric proportional hazards model.

Before proceeding further, it is useful to place this article in the context of previous work on nonignorable drop-out. Recent years have brought an explosive growth of literature in this area. This is reflective of the increasing recognition that subjects may drop out of a longitudinal study because of factors related either directly or indirectly to the outcome under investigation. The literature can be divided into likelihood-based and nonlikelihood-based approaches. In the likelihood framework, full parametric specification of the joint distribution of outcomes and the nonresponse mechanism is required. Hogan and Laird (1997a) and Little (1995) have provided reviews of much of this literature, including the work of DeGruttola and Tu (1994), Diggle and Kenward (1994), Fitzmaurice, Clifford, and Heath (1996), Fitzmaurice, Laird, and Zahner (1996), Fitzmaurice, Molenberghs, and Lipsitz (1995), Hogan and Laird (1997b), Laird (1988), Little (1993a,b), Mori, Woodworth, and Woolson (1992), Schluchter (1992), Self and Pawitan (1992), Tsiatis, DeGruttola, and Wulfsohn (1994), Wu and Bailey (1988, 1990), Wu and Carroll (1988). In the nonlikelihood approach considered by Robins (1997), Robins, Rotnitzky, and Zhao (1995), Rotnitzky and Robins (1997), and Rotnitzky et al. (1998), the joint distribution of the outcomes is assumed to follow a nonparametric or semiparametric model, whereas the nonresponse mechanism is assumed to follow a parametric model. The current work extends the nonlikelihood approach by allowing for semiparametric nonresponse mechanisms.

In the next section we describe our data structure and define and motivate models for these data. We discuss issues of identifiability of parameters in these models, which lead

Daniel O. Scharfstein is Assistant Professor of Biostatistics, Johns Hopkins School of Hygiene and Public Health, Baltimore, MD 21205. Andrea Rotnitzky is Associate Professor of Biostatistics and James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. This research was partially supported by National Institute of Health grants 1-R29-GM48704-0, 5R01A132475-07, R01CA74112, 1-R01-MH56639-01A1, R01-HD-38209-01, and 1-R01-DA10184-01A2. The authors wish to thank Victor DeGruttola for helpful discussions.

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

us into a exposition of our philosophy of sensitivity analysis. We also set the stage for the remainder of the article.

2. DATA, MODELS, IDENTIFIABILITY, AND PHILOSOPHY OF SENSITIVITY ANALYSIS

2.1 Identifiability

We assume that we observe n iid copies, $\{O_i = (Q_i, \Delta_i, \Delta_i Y_i, \bar{\mathbf{V}}_i(Q_i))\colon i = 1, \ldots, n\}$, of

$$O = (Q, \Delta, \Delta Y, \bar{\mathbf{V}}(Q)),$$

where Q is time to drop-out, Y is the outcome of interest measured at the fixed nonrandom end-to-follow-up time $T, \Delta = I(Q \geq T)$ is the drop-out indicator, and $\bar{\mathbf{V}}(t) = \{\mathbf{V}(u)\colon 0 \leq u \leq t\}$ is the history of all other variables that would be recorded through time t in the absence of drop-out. Note that Y is observed if and only if $\Delta = 1$. For notational convenience, for subjects who do not drop out, we set the drop-out time Q equal to the end of follow-up time T.

The goal of this article is to consider inference about a smooth functional of the marginal distribution of Y using the observed data $\{O_i, i = 1, \ldots, n\}$. For concreteness, we concentrate on inference about the mean μ_0 of Y, although we briefly consider the median in Section 7.1.

Consider model A, in which we assume that the conditional hazard of Q, given the data $(\bar{\mathbf{V}}(T), Y)$ that would be observed in the absence of drop-out, follows a stratified Cox proportional hazards model of the form

$$\lambda_Q(t|\bar{\mathbf{V}}(T), Y) = \lambda_0(t|\bar{\mathbf{V}}(t)) \exp(\alpha_0 Y), \qquad (1)$$

where $\lambda_Q(t|\bar{\mathbf{V}}(T), Y) = \lim_{h \to 0} \Pr[t \leq Q < t + h|\bar{\mathbf{V}}(T), Y, Q \geq t]/h, \lambda_0(t|\bar{\mathbf{V}}(t))$ is an unrestricted positive function, and α_0 is an unknown parameter. Equation (1) states that the hazard of drop-out at time t depends in an arbitrary and unknown way on the observed past $\bar{\mathbf{V}}(t)$, but depends on the possibly unobserved future only through the term $\exp(\alpha_0 Y)$. When $\alpha_0 = 0$, drop-out at time t is conditionally independent of the possibly unobserved outcome Y given the observed past $\bar{\mathbf{V}}(t)$. Thus $\alpha_0 = 0$ corresponds to the assumption that the data are coarsened at random (CAR) as defined by Heitjan and Rubin (1991). Robins and Rotnitzky (1992) previously studied inference in model A when $\alpha_0 = 0$. This article extends their work by allowing for nonzero α_0. The sensitivity analysis philosophy that we adopt herein is motivated by the following identification theorem, whose proof is given in Appendix A.

Theorem 1. Under the regularity conditions given in Appendix A, (a) in model A, neither α_0 nor the distribution of $(\bar{\mathbf{V}}(T), Y)$ is identified, and (b) for each law F_O of the observed data O and each value of α_0, there exists a unique $\lambda_0(t|\bar{\mathbf{V}}(t))$ and a unique joint law, say $F_{\bar{\mathbf{V}}(T),Y}$, of $(\bar{\mathbf{V}}(T), Y)$ such that F_O is the marginal distribution of O under the law $F_{\bar{\mathbf{V}}(T),Y,Q}$ for $(\bar{\mathbf{V}}(T), Y, Q)$ determined by $F_{\bar{\mathbf{V}}(T),Y}$ and (1).

Consider model $A(\alpha_0)$, which differs from model A only in that α_0 *is assumed known to the data analyst.* Theorem 1(b) implies that model $A(\alpha_0)$ is nonparametric (i.e., saturated) for F_O in the sense that it places no restrictions on the joint law of the observed data. That is, model $A(\alpha_0)$ fits the data perfectly and cannot be rejected by any statistical test. The theorem also implies that model $A(\alpha_0)$ is just identified in the sense that the joint distribution of $(\bar{\mathbf{V}}(T), Y, Q)$ is uniquely determined by the law of the observables. We thus refer to model $A(\alpha_0)$ as a "nonparametric (just) identified" (NPI) model. In (A.1a)–(A.1c) of Appendix A we give an explicit characterization of the map from (F_O, α_0) to the law $F_{\bar{\mathbf{V}}(T),Y,Q}$.

It follows that under model $A(\alpha_0)$, for any given law F_O of O, we can plot the mean of Y, say $\mu(\alpha_0)$, as a function of the nonidentified selection bias parameter α_0. In practice, the law F_O is unknown, but it can be estimated from the observed data. Thus we can replace $\mu(\alpha_0)$ by an estimator $\hat{\mu}(\alpha_0)$ and provide a confidence interval for $\mu(\alpha_0)$ that will be guaranteed to asymptotically cover $\mu(\alpha_0)$ at its nominal rate. Our estimation procedure is described in Section 3. We now illustrate our approach with a concrete example.

2.2 Example: ACTG 175

ACTG 175 was a randomized, double-blind clinical trial designed to evaluate nucleoside monotherapy (zidovudine or didanosine) versus combination therapy (zidovudine/didanosine or zidovudine/zalcitabine) in HIV-1 infected individuals with CD4 cell counts of 200–500/mm^3. Specifically, 2,467 subjects were randomized to one of four treatment arms: (1) zidovudine 200 mg three times daily (AZT); (2) zidovudine 200 mg three times daily plus didanosine 200 mg twice daily (AZT + ddI); (3) zidovudine 200 mg three times daily plus zalcitabine .75 mg daily (AZT + ddC); and (4) didanosine 200 mg twice daily (ddI). Enrollment began in December 1991 and was closed in October 1992. CD4 counts were obtained at baseline and again at weeks 8, 20, 32, 44, and 56 (Hammer et al. 1996).

One goal of the investigators was to compare the four treatment arm–specific mean CD4 counts at week 56 had (possibly contrary to fact) all subjects complied with their assigned therapy through that week. This goal differs from that of an intent-to-treat analysis, which aims to compare treatment arm–specific means of subjects as randomized, regardless of compliance. Therefore, for the purpose of our analyses, subjects were considered to be drop-outs if they died or were lost to follow-up prior to week 56, if they missed their 56 week clinic visit, or if they were observed to discontinue their assigned therapy prior to week 56. Drop-outs varied from 26.5 to 36% in the four arms. An approximate time to drop-out was available for these subjects. Note that, as is common in randomized trials, our interest is in the unconditional mean of CD4 count Y at week 56 rather than the conditional mean of Y given $\bar{\mathbf{V}}(t)$. The conditional mean of Y given baseline covariates $\mathbf{V}(0)$ might be of interest for "subset" analyses, which are not considered in this article. Figure 1 presents a sensitivity analysis based on model $A(\alpha_0)$ in which we took $\bar{\mathbf{V}}(t)$ to be the time-independent indicator of whether the subject was an IV drug user at baseline. In Figure 1 we show the estimated means along with 95% confidence intervals for

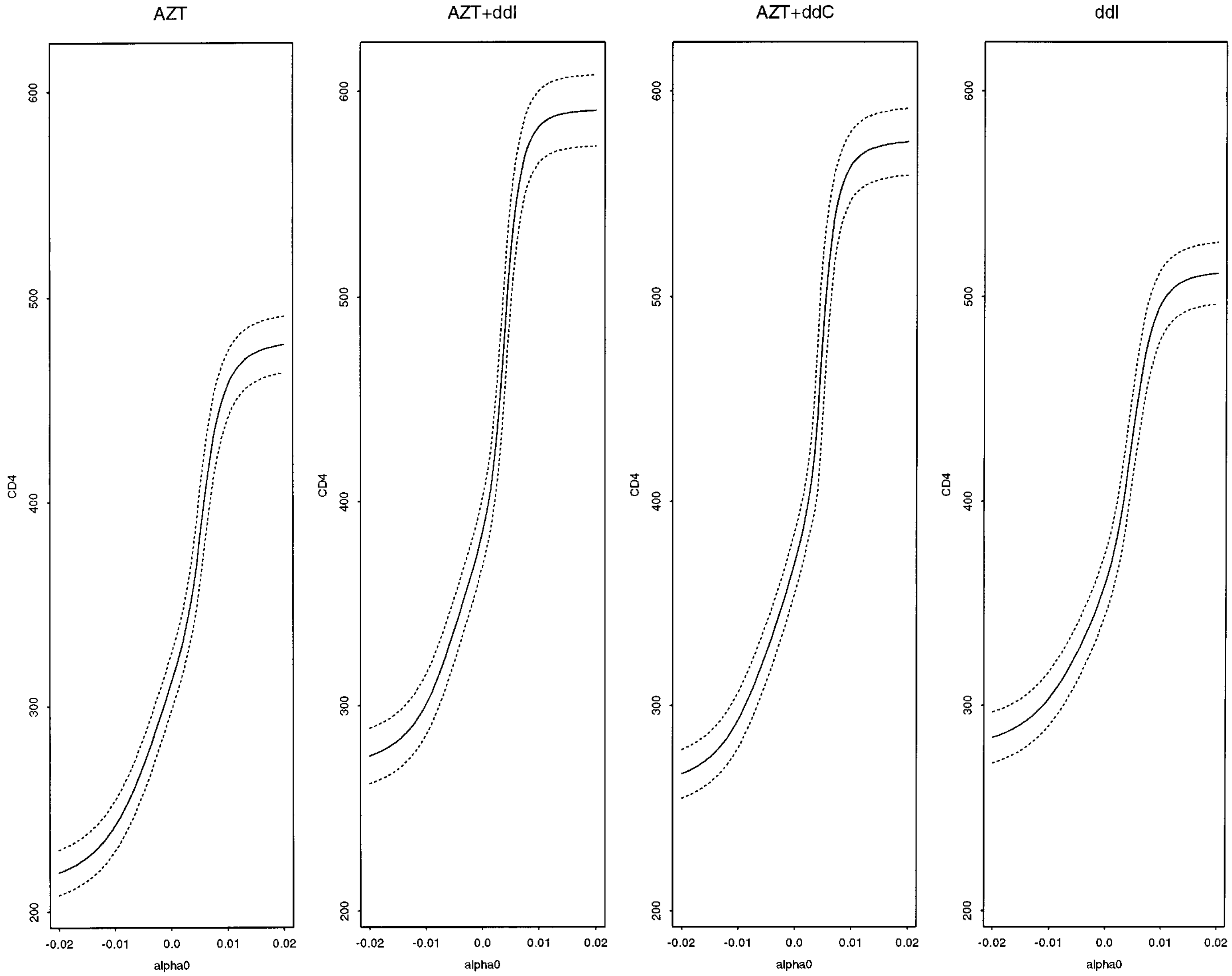

Figure 1. Treatment-Specific Predicted Means and 95% Confidence Intervals for Varying α_0's in Model A(α_0) with IV Drug User Status as the Time-independent Regressor.

α_0's ranging from $-.02$ to $.02$ for each of the four treatment groups. Between-treatment group comparisons are reported in Section 6. In Figure 1 α_0 is interpreted as the log hazard ratio for drop-out between subjects with the same baseline IV drug user status, but who differ by 1 in CD4 count at week 56. Setting $\alpha_0 > 0$ (< 0) is tantamount to assuming that among subjects with the same IV drug user status, those with higher (lower) CD4 counts at 56 weeks are more likely to be drop-outs than those with lower (higher) CD4 counts. For example, setting $\alpha_0 = .01$ specifies that at each time t an IV drug user with a 200 CD4 count at 56 weeks has a drop-out hazard 2.7 times that of an IV drug user with a 100 CD4 count at 56 weeks and a drop-out hazard $(2.7)^{3/2}$ times that of a drug user with a CD4 count of 50 at 56 weeks. As expected, for each treatment group, the estimated means increased monotonically with α_0.

2.3 A Philosophy of Sensitivity Analysis

The reader should not be discouraged that we only provide a sensitivity analysis for the mean of Y. Because the parameter α_0 represents the magnitude of selection bias due to unmeasured factors, it would not be desirable or scientifically reasonable for α_0 to be identified from the data in the absence of further knowledge of these factors. Our model $A(\alpha_0)$ formalizes this desiderata; we cannot identify the magnitude of selection bias, but we can identify the law of Y, and in particular its mean, as a function of the selection bias parameter. Because the data contain no independent evidence about α_0, final substantive conclusions would depend on which values of α_0 are considered plausible by relevant subject matter experts. In Appendix A we prove that Theorem 1 remains true if we replace $\alpha_0 Y$ in (1) by any other fixed function $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ of $t, \alpha_0, \bar{\mathbf{V}}(T)$, and Y that satisfies the regularity conditions of Appendix A. Thus there will never be any data evidence that can determine either the magnitude of α_0 or the functional form $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ of the selection bias process. It follows that one may wish to repeat the preceding sensitivity analysis with $\alpha_0 Y$ replaced by other functional forms $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ satisfying $r(t, 0; \bar{\mathbf{V}}(T), Y) = 0$, so $\alpha_0 = 0$ continues to imply CAR. Note that the substantive meaning of the magnitude of α_0 depends on the functional form chosen for $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$. Sensitivity analyses based on other NPI models have been considered in the missing-data literature by Baker, Rosenberger, and DerSimonian (1992), Nordheim (1984), Robins (1997), and Rotnitzky et al. (1998) and in the competing risks literature by Klein and Moeschberger (1988), Slud and Rubinstein (1983), and Zheng and Klein (1994, 1995). With the exception of those of Robins (1997) and Rotnitzky et al. (1998), these NPI models do not allow for time-dependent processes $\bar{\mathbf{V}}(t)$.

It is important to note that it would be possible to jointly identify the nonignorable selection bias parameter α_0 and the mean μ_0 of Y (as well as estimate them at $n^{1/2}$-rates) if, in addition to (1), we specified that either $F_{Y,\bar{\mathbf{V}}(T)}$ or $\lambda_0(t|\bar{\mathbf{V}}(t))$ followed particular parametric models. However, one would rarely, if ever, have such firm prior knowledge of the functional form of either $F_{Y,\bar{\mathbf{V}}(T)}$ or $\lambda_0(t|\bar{\mathbf{V}}(t))$ so that such parametric restrictions should be used to identify the parameters of scientific interest. Little (1985) and Little and Rubin (1987) expressed similar sentiments. In our opinion, it is preferable that μ_0 and α_0 not be jointly identified from the data in the absence of additional well-supported substantive knowledge. This position is in line with the adage: "It's not what you don't know that hurts you; it's the things you think you know, but don't." To paraphrase Freedman, Rothenberg, and Sutch (1984), identifying α_0 and μ_0 by specifying parametric models will increase the stock of things that we think we know, but do not.

Clearly the biggest challenge in conducting such a sensitivity analysis is the choice of one or more sensible parameterized selection bias functions $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ whose interpretation can be communicated to relevant subject matter experts with sufficient clarity so that they can provide a plausible range for the parameter α_0, including its magnitude and direction. Hard as this challenge may sound, we believe it to be a worthwhile exercise when compared to the alternative of identifying α_0 and μ_0 based on poorly motivated parametric functional forms and/or distributional shape restrictions.

When a decision is required (e.g., whether a drug should be licensed based on the study results), a drawback of sensitivity analysis is that it produces a range of answers rather than a single answer. In this case it would be reasonable to place a prior distribution on the nonidentified selection bias parameter α_0, and also on the functional form of $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$. Robins, Rotnitzky, and Scharfstein (1999, sec. 11) provided details of this approach, although the discussion there is restricted to a rather simple setting because of unsolved technical problems with implementing nonparametric Bayesian procedures. Even if one wished to summarize inferences by Bayesian averaging over possible values of α_0, we recommend that one also publish the results of the sensitivity analysis itself, to make the reader aware of how inferences about μ_0 vary with α_0. In this sense we regard a sensitivity analysis as useful "preprocessing" for any full Bayesian analysis that places prior distributions on α_0 and the other parameters.

2.4 The Curse of Dimensionality

Theorem 1 guarantees that both the law of $(\bar{\mathbf{V}}(T), Y)$ and $\lambda_0(t|\bar{\mathbf{V}}(t))$ are nonparametrically identified under model $A(\alpha_0)$. Furthermore, in Section B.1.1 of Appendix B we prove that the asymptotic semiparametric information bound for $n^{1/2}$-consistent estimators of μ_0 in model $A(\alpha_0)$ is nonzero. Nonetheless, neither Theorem 1 nor the positive information bound guarantees that we can construct estimators of μ_0 that perform well in the moderate-sized samples found in practice. In fact, when $\bar{\mathbf{V}}(t)$ is high dimensional and the sample size is moderate (say, less than 1,000), then, due to the curse of dimensionality, there is no estimator of μ_0 that has, under all laws allowed by model $A(\alpha_0)$, an approximately normal sampling distribution centered near μ_0 with variance sufficiently small to be of substantive interest (Robins and Ritov 1997). This reflects the fact that to estimate μ_0 under model $A(\alpha_0)$, it is necessary to use multivariate nonparametric smoothing techniques, which would require impractically large samples when $\bar{\mathbf{V}}(t)$ is high dimensional.

We regard the process $\bar{\mathbf{V}}(t)$ as high dimensional if (a) for each t, the vector $\mathbf{V}(t)$ has two or more continuous components or many discrete components, or (b) $\bar{\mathbf{V}}(t)$ jumps at many different times. In such cases we consider model B, in which we assume that (1) holds and

$$\lambda_0(t|\bar{\mathbf{V}}(t)) = \lambda_0(t) \exp(\boldsymbol{\gamma}_0' \mathbf{W}(t)), \tag{2}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\mathbf{W}(t) = \mathbf{w}(t, \bar{\mathbf{V}}(t)), \mathbf{w}(\cdot, \cdot)$ is a known function that maps $(t, \bar{\mathbf{V}}(t))$ to R^q, and $\boldsymbol{\gamma}_0$ is a q-dimensional unknown parameter. In Section 4 we show that inference about μ_0 under model $B(\alpha_0)$ (i.e., model B with α_0 assumed known) does not require high-dimensional smoothing. Thus we can leave the baseline hazard $\lambda_0(t)$ unrestricted and still obtain well-behaved estimates of μ_0.

Theorem 1 does not hold for model B. As a consequence of the restriction on the functional form of $\lambda_0(t|\bar{\mathbf{V}}(t))$ imposed by (2), μ_0 and α_0 are often jointly identified. But if we choose the dimension of $\mathbf{W}(t)$ in (2) moderately large, to preserve some robustness to misspecification, then there would be generally little independent information about α_0 and μ_0, and thus their joint estimation would require very large sample sizes. Thus we continue to recommend that one regard α_0 as fixed and known when estimating μ_0 and vary α_0 in a sensitivity analysis. As this model $B(\alpha_0)$ is no longer a nonparametric model for the distribution of the observed data, it can in principle be subjected to a goodness-of-fit test. In conducting a sensitivity analysis, we would like to choose the dimension of $\mathbf{W}(t)$ in (2) large enough so that any goodness-of-fit test will have little power to reject model $B(\alpha_0)$, but choose the dimension small enough so that the estimators described in Section 4 have a nearly normal sampling distribution with variance small enough to be of substantive use to subject matter experts. It is not clear that both of these competing criteria can always be met. Clearly, the choice of the dimension of $\mathbf{W}(t)$ will depend on the size of the dataset and on the precision required by the experts. Furthermore, because different models $B(\alpha_0)$

associated with different choices for the functional form of $\mathbf{W}(t) = \mathbf{w}(t, \bar{\mathbf{V}}(t))$ cannot be easily distinguished based on a goodness-of-fit test and may lead to quite different inferences for μ_0, it would be best to repeat a sensitivity analysis a number of times, varying not only the functional form of the nonidentified selection bias function $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$, but also the functional form of $\mathbf{w}(t, \bar{\mathbf{v}}(t))$ in (2).

In Sections 3 and 4 we show how to estimate μ_0 in models $A(\alpha_0)$ and $B(\alpha_0)$. These estimation procedures are a special case of a general theory of inference in settings in which the full data (i.e., $\bar{\mathbf{V}}(T)$ and Y) and the drop-out mechanism (i.e., Q) follow arbitrary semiparametric models with distinct parameters. For ease of presentation, we describe this general theory and then give specific applications to the models considered in this article in Appendix B. In Section 5 we present the results of two simulation studies that evaluate the performance of our estimation procedures in moderate-sized samples. In Section 6 we perform a sensitivity analysis of the ACTG 175 dataset using our two models. In Section 7 we describe settings in which our method breaks down and offer alternative methods appropriate for these settings. We denote the final section to a discussion.

3. ESTIMATION IN MODEL $A(\alpha_o)$

To motivate our estimation method, suppose first that $\lambda_0(t|\bar{\mathbf{V}}(t))$ were known. Let $\Lambda_0(t|\bar{\mathbf{V}}(t)) = \int_0^t \lambda_0(u|\bar{\mathbf{V}}(u))\,du$ denote the cumulative conditional baseline hazard. Then we could estimate μ_0 by $\tilde{\mu}(b)$ solving

$$\sum_{i=1}^{n} h(O_i; \mu, \Lambda_0; b) = 0,$$

where $b = b(\bar{\mathbf{v}}(t), t; \mu)$ is a function specified by the data analyst and

$$\begin{aligned} &h(O; \mu, \Lambda_0; b) \\ &\quad = \frac{\Delta}{\pi(\bar{\mathbf{V}}(T), Y)} \\ &\qquad \times (Y - \mu - E[(1-\Delta) b(\bar{\mathbf{V}}(Q), Q; \mu)|\bar{\mathbf{V}}(T), Y]) \\ &\qquad + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q; \mu), \end{aligned} \tag{3}$$

with $\pi(\bar{\mathbf{V}}(T),\ Y) = \Pr[\Delta = 1|\bar{\mathbf{V}}(T),\ Y]$. By (1), we have $\Pr[\Delta = 1|\bar{\mathbf{V}}(T),\ Y] = S(T|\bar{\mathbf{V}}(T),\ Y)$, where $S(t|\bar{\mathbf{V}}(T), Y) = \exp(-\Lambda_0(t|\bar{\mathbf{V}}(t))\exp(\alpha_0 Y))$. Furthermore, the conditional expectation in (3) can be explicitly evaluated as $\int_{0\_}^{T} b(\bar{\mathbf{V}}(t),\ t;\ \mu)\exp(-\Lambda_0(t|\bar{\mathbf{V}}(t))\exp(\alpha_0 Y))\exp(\alpha_0 Y)$ $\lambda_0(t|\bar{\mathbf{V}}(t))\,dt$. In the special case in which $b(\bar{\mathbf{v}}(t), t; \mu)$ is chosen to be identically 0, we refer to $\tilde{\mu}(b)$ as an inverse probability of censoring weighted (IPCW) estimator. This is a generalization of the Horvitz–Thompson (Horvitz and Thompson 1952) estimator used in the sample survey literature. When $b(\bar{\mathbf{v}}(t), t; \mu)$ is nonzero, we refer to $\tilde{\mu}(b)$ as an augmented IPCW (AIPCW) estimator.

The regularity condition 2 of Appendix A guarantees that $\pi(\bar{\mathbf{V}}(T), Y) > 0$ with probability 1. Thus $E[\Delta/\pi(\bar{\mathbf{V}}(T), Y)|\bar{\mathbf{V}}(T), Y] = 1$, and hence $E[h(O; \mu_0, \Lambda_0; b)] = 0$ for any function b. Using standard Taylor series arguments, it can be shown that $n^{1/2}(\tilde{\mu}(b) - \mu_0)$ is asymptotically normal with mean 0 and asymptotic variance $\tau(b)^{-2}$ var$[h(O; \mu_0, \Lambda_0; b)]$ where $\tau(b) = \partial E[h(O; \mu, \Lambda_0; b)]/\partial\mu|_{\mu=\mu_0}$. Given a consistent estimator $\tilde{\tau}(b)$ of $\tau(b)$, the asymptotic variance can be consistently estimated by $\tilde{\tau}(b)^{-2} n^{-1} \sum_{i=1}^{n} h(O_i; \tilde{\mu}(b), \Lambda_0; b)^2$. When the functional μ_0 is the mean of $Y, \tau(b) = 1$, and $\tilde{\tau}(b)$ can be taken to be 1 as well.

Because in fact $\lambda_0(u|\bar{\mathbf{V}}(u))$, and thus $\pi(\bar{\mathbf{V}}(T), Y)$ and the conditional expectations in (3) are unknown, we consider estimators $\hat{\mu}(b)$ solving

$$\sum_{i=1}^{n} h(O_i; \mu, \hat{\Lambda}; b) = 0,$$

where

$$\begin{aligned} &h(O; \mu, \hat{\Lambda}; b) \\ &\quad = \frac{\Delta}{\hat{\pi}(\bar{\mathbf{V}}(T), Y)} \\ &\qquad \times (Y - \mu - \hat{E}[(1-\Delta) b(\bar{\mathbf{V}}(Q), Q; \mu)|\bar{\mathbf{V}}(T), Y]) \\ &\qquad + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q; \mu), \end{aligned}$$

$\hat{\pi}(\bar{\mathbf{V}}(T), Y)$ is equal to $\exp(-\hat{\Lambda}(T|\bar{\mathbf{V}}(T))\exp(\alpha_0 Y))$, $\hat{E}[(1-\Delta) b(\bar{\mathbf{V}}(Q),\ Q;\ \mu)|\bar{\mathbf{V}}(T), Y]$ is equal to $\int_0^T b(\bar{\mathbf{V}}(t),\ t;\ \mu)$ $\exp(-\hat{\Lambda}(t|\bar{\mathbf{V}}(t))\exp(\alpha_0 Y))\exp(\alpha_0 Y)\,d\hat{\Lambda}(t|\bar{\mathbf{V}}(t))$, and $\hat{\Lambda}(t|$ $\bar{\mathbf{V}}(t))$ is the estimate of the cumulative baseline hazard $\Lambda_0(t|\bar{\mathbf{V}}(t))$ described later.

Unfortunately, due to the curse of dimensionality, nonparametric estimation of $\Lambda_0(t|\bar{\mathbf{V}}(t))$ is not feasible when $\mathbf{V}(t)$ has multiple continuous components or the process $\bar{\mathbf{V}}(t)$ jumps at many different times. In the remainder of this section we consider the special case in which $\bar{\mathbf{V}}(t)$ is time independent so that $\bar{\mathbf{V}}(t) = \mathbf{V}$ for all t.

If $\mathbf{V}$ is discrete, then $\hat{\Lambda}(t|\mathbf{V})$ is estimated separately within each level of $\mathbf{V}$. If $\mathbf{V}$ is univariate and continuous, then $\hat{\Lambda}(t|\mathbf{V})$ is nonparametrically estimated by a "histogram" estimator that places subjects with similar values of $\mathbf{V}$ into a common bin and constructs estimators $\hat{\Lambda}(t|\mathbf{V})$ separately for each bin. Suppose that $\mathbf{V}$ is discrete or has been discretized by grouping into a finite number of "bins." If we were always able to observe Y, then we could simply partition the sample into groups based on the value of $\mathbf{V}$ and estimate the cumulative baseline hazard separately within each group using the Nelson–Aalen estimator with censoring times as the jump times (Andersen, Borgan, Gill, and Keiding 1993). That is, we could estimate $\Lambda_0\ (t|\mathbf{V} = \mathbf{v})$ by

$$\begin{aligned} &\tilde{\Lambda}(t|\mathbf{V} = \mathbf{v}) \\ &\quad = \int_0^t \left(\frac{1}{n_{\mathbf{v}}} \sum_{i=1}^{n} I(\mathbf{V}_i = \mathbf{v}) \exp(\alpha_0 Y_i) I(Q_i \geq u) \right)^{-1} \\ &\qquad\qquad \times \left(\frac{1}{n_{\mathbf{v}}} \sum_{i=1}^{n} dN_i^{\mathbf{v}}(u) \right), \end{aligned} \tag{4}$$

where $N_i^{\mathbf{v}}(u) = I(\mathbf{V}_i = \mathbf{v}, Q_i \le u, \Delta_i = 0)$ and $n_{\mathbf{v}} = \sum_{i=1}^n I(\mathbf{V}_i = \mathbf{v})$. Because Y is not always observed, we need to modify the foregoing estimator. The integrand of (4) is not observable. So we would like to replace it with an observable quantity that has the same probability limit. The key observation is that $E[\Delta I(\mathbf{V} = \mathbf{v}) S(u|\mathbf{V} = \mathbf{v}, Y)/S(T|\mathbf{V} = \mathbf{v}, Y)|\mathbf{V} = \mathbf{v}, Y, Q \ge u] = 1$, so that by a uniform large of large numbers,

$$\sup_{u\in[0,T]} \left| \frac{1}{n_{\mathbf{v}}} \sum_{i=1}^n I(\mathbf{V}_i = \mathbf{v}) \exp(\alpha_0 Y_i) I(Q_i \ge u) - \frac{1}{n_{\mathbf{v}}} \sum_{i=1}^n \frac{\Delta_i I(\mathbf{V}_i = \mathbf{v}) \exp(\alpha_0 Y_i) I(Q_i \ge u)}{\exp\{-\exp(\alpha_0 Y_i)(\Lambda_0(T|\mathbf{V}_i = \mathbf{v}) - \Lambda_0(u|\mathbf{V}_i = \mathbf{v}))\}} \right| \xrightarrow{P} 0.$$

The latter quantity within the absolute value depends on $\Lambda_0(\cdot|\mathbf{V} = \mathbf{v})$, but we can substitute $\hat{\Lambda}(\cdot|\mathbf{V} = \mathbf{v})$ to define the recursive estimator

$$\hat{\Lambda}(t|\mathbf{V} = \mathbf{v}) = \int_0^t \left(\frac{1}{n_{\mathbf{v}}} \sum_{i=1}^n \frac{\Delta_i I(\mathbf{V}_i = \mathbf{v}) \exp(\alpha_0 Y_i) I(Q_i \ge u)}{\exp(-\exp(\alpha_0 Y_i)(\hat{\Lambda}(T|\mathbf{V}_i = \mathbf{v}) - \hat{\Lambda}(u|\mathbf{V}_i = \mathbf{v})))} \right)^{-1} \times \left(\frac{1}{n_{\mathbf{v}}} \sum_{i=1}^n dN_i^{\mathbf{v}}(u) \right).$$

We can obtain an explicit solution for $\hat{\Lambda}(t|\mathbf{V} = \mathbf{v})$ as follows. First, note that $\hat{\Lambda}(t|\mathbf{V} = \mathbf{v})$ is a step function with jumps at each of the unique censoring times in the group with $\mathbf{V} = \mathbf{v}$. Thus we need only compute the jump sizes. Henceforth, let $Q_{(1)}^{\mathbf{v}}, < \ldots, < Q_{(k_{\mathbf{v}})}^{\mathbf{v}}$ denote these unique times. Let $c_k^{\mathbf{v}}$ denote the number of subjects who are censored at $Q_{(k)}^{\mathbf{v}}, k = 1, \ldots, k_{\mathbf{v}}$. Note that when, as we have assumed, Q has a continuous distribution function, $c_k^{\mathbf{v}}$ will only take value 1. Thus we can write $\hat{\Lambda}(t|\mathbf{V} = \mathbf{v}) = \sum_{k=1}^{k_{\mathbf{v}}} \hat{\lambda}_k^{\mathbf{v}} I(Q_{(k)}^{\mathbf{v}} \le t)$, where the jump size $\hat{\lambda}_k^{\mathbf{v}}$ is found by the following procedure:

1. $\hat{\lambda}_{k_{\mathbf{v}}}^{\mathbf{v}} = (\sum_{i=1}^n \Delta_i I(\mathbf{V}_i = \mathbf{v}) \exp(\alpha_0 Y_i))^{-1} c_{k_{\mathbf{v}}}^{\mathbf{v}}$.
2. For $k = k_{\mathbf{v}} - 1, \ldots, 1$, sequentially compute

$$\hat{\lambda}_k^{\mathbf{v}} = \left(\sum_{i=1}^n \frac{\Delta_i I(\mathbf{V}_i = \mathbf{v}) \exp(\alpha_0 Y_i)}{\exp(-\exp(\alpha_0 Y_i) \sum_{j=k+1}^{k_{\mathbf{v}}} \hat{\lambda}_j^{\mathbf{v}})} \right)^{-1} c_k^{\mathbf{v}}.$$

Under regularity conditions, we would expect $n^{1/2}\{\hat{\Lambda}(\cdot|\mathbf{V} = \mathbf{v}) - \Lambda_0(\cdot|\mathbf{V} = \mathbf{v})\}$ to converge to a Gaussian process. Hence we consistently estimate the survivor function of Q given $\mathbf{V} = \mathbf{v}$ and Y by $\hat{S}(t|\mathbf{V} = \mathbf{v}, Y) = \exp(-\hat{\Lambda}(t|\mathbf{V} = \mathbf{v}) \exp(\alpha_0 Y))$.

If $\mathbf{V}$ is discrete or $\mathbf{V}$ is univariate and continuous, $\lambda_0(t|\mathbf{V})$ is smooth as a function of $\mathbf{V}$, and the binwidth is decreased with increasing sample size at an appropriate rate, then, under some additional mild regularity conditions $\hat{\mu}(b)$ should be a regular and asymptotically linear (RAL) estimator of μ_0 with influence function $d(O; b)$. Recall that an estimator $\hat{\mu}(b)$ is asymptotically linear with influence function $d(O; b)$ if $n^{1/2}(\hat{\mu}(b) - \mu_0) = n^{-1/2} \sum_{i=1}^n d(O_i; b) + o_p(1)$, where $E[d(O; b)] = 0, 0 < E[d(O; b)^2] < \infty$, and $o_p(1)$ refers to a random variable converging to 0 in probability. An estimator $\hat{\mu}(b)$ of μ_0 is regular in a semiparametric model if its convergence to μ_0 is locally uniform. (See Bickel, Klassen, Ritov, and Wellner 1993 for a more precise definition.) Regularity is a technical condition imposed to prohibit superefficient estimators. In fact, even a fully parametric model will have nonregular estimators whose asymptotic variance is less than the Cramer–Rao variance bound. If $\hat{\mu}(b)$ is asymptotically linear, then $n^{1/2}(\hat{\mu}(b) - \mu_0)$ is asymptotically normal with mean 0 and variance $E[d(O; b)^2]$. Further, any two asymptotically linear estimators with the same influence function are asymptotically equivalent in the sense that $n^{1/2}$ times their difference converges to 0 in probability. In general, the influence function $d(O; b)$ of $\hat{\mu}(b)$ will not be equal to the influence function $-\tau(b)^{-1} h(O; \mu_0, \Lambda_0; b)$ of $\tilde{\mu}(b)$, and the asymptotic variance $E[d(O; b)^2]$ will not be consistently estimated by

$$n^{-1} \tilde{\tau}(b)^{-2} \sum_{i=1}^n h(O_i; \hat{\mu}(b), \hat{\Lambda}; b)^2. \tag{5}$$

This is because for most choices of b, estimation of $\Lambda_0(t|\mathbf{V})$ contributes a term to the asymptotic variance of $\hat{\mu}(b)$, in which case the asymptotic variance $\tau(b)^{-2} E[h(O; \mu_0, \Lambda_0; b)^2]$ of $\tilde{\mu}(b)$ cannot be the same as $E[d(O; b)^2]$.

In fact, regardless of the choice of the function b, all estimators $\hat{\mu}(b)$ will have the same influence function with asymptotic variance equal to the semiparametric variance bound for estimators of μ_0 in model $A(\alpha_0)$. This follows from the fact that by Theorem 1, model $A(\alpha_0)$ is a nonparametric model for the observed data O, and Bickel et al. (1993) proved that for any nonparametric model, all RAL estimators of any functional of F_O (such as μ_0) have the same influence function. Thus if we can find a function b^* for which estimation of $\Lambda_0(t|\mathbf{V})$ does not contribute to the asymptotic variance of $\hat{\mu}(b)$, then for all $b, \hat{\mu}(b)$ will have influence function $-\tau(b^*)^{-1} h(O; \mu_0, \Lambda_0; b^*)$. For $b \ne b^*$, the asymptotic variance of $\hat{\mu}(b)$ will be consistently estimated not by (5), but rather by

$$n^{-1} \tilde{\tau}(b^*)^{-2} \sum_{i=1}^n h(O_i; \hat{\mu}(b), \hat{\Lambda}; b^*)^2. \tag{6}$$

We now present a heuristic approach to finding b^*. Our estimating function $h(O; \mu, \Lambda_0; b)$ depends on the unrestricted infinite-dimensional nuisance parameter $\Lambda_0(t|\mathbf{V})$. Estimation of $\Lambda_0(t|\mathbf{V})$ does not contribute to the asymptotic variance of $\hat{\mu}(b)$ when $\Lambda_0(t|\mathbf{V})$ is unrestricted if and only if the same is true when $\Lambda_0(t|\mathbf{V})$ follows any arbitrary correctly specified parametric submodel. But if $\Lambda_0(t|\mathbf{V})$, or, equivalently $\lambda_0(t|\mathbf{V})$, had a known parametric form indexed by η with true value η_0 and estimated value $\hat{\eta}$, then we could write $h(O; \mu, \Lambda_0; b)$ as $h^{\dagger}(O; \mu, \eta_0; b)$ and expand $h^{\dagger}(O; \mu, \hat{\eta}; b)$ around η_0 to derive the asymptotic variance of $\hat{\mu}(b)$. Inspection of the Taylor expansion terms would reveal that a necessary and sufficient condition for the estimation of η not to affect the asymptotic variance of $\hat{\mu}(b)$ is that $n^{-1}\sum_i \partial h^{\dagger}(O_i;\mu_0,\eta_0;b)/\partial\eta$ converge to 0 in probability, or, equivalently, that $E[\partial h^{\dagger}(O;\mu_0,\eta_0;b)/\partial\eta]=0$. But it can be shown (Newey 1990) that $E[\partial h^{\dagger}(O;\mu_0,\eta_0;b)/\partial\eta] = E[h(O;\mu_0,\Lambda_0;b)S_\eta]$, where $S_\eta = \partial \ln \mathcal{L}(\mu_0,\eta_0;O)/\partial\eta$ is the derivative of the observed-data log-likelihood $\ln\mathcal{L}(\mu_0,\eta;O)$ for a single subject with respect to η. Thus we conclude that estimation of $\Lambda_0(t|\mathbf{V})$ will not contribute to the asymptotic variance only for those choices of b such that $h(O;\mu_0,\Lambda_0;b)$ is uncorrelated with the scores S_η of all parametric submodels for $\lambda_0(t|\mathbf{V})$. In Appendix B we show that there exists one and only one such function b^*, given by

$$\begin{aligned} b^*(\mathbf{V},t;\mu) &= b^*(\mathbf{V};\mu) \\ &= E[(Y-\mu)\exp(\alpha_0 Y)|\mathbf{V}]/E[\exp(\alpha_0 Y)|\mathbf{V}]. \end{aligned} \tag{7}$$

The function $b^*(\mathbf{V};\mu)$ is not available for data analysis, because it depends on the unknown conditional expectations in (7). However, using the arguments of Robins, Mark, and Newey (1992), it can be shown that if $\hat{b}^*(\mathbf{v};\mu)$ is a consistent estimator of $b^*(\mathbf{v};\mu)$ for each $\mathbf{v}$, then $\hat{\mu}(b^*)$ and $\hat{\mu}(\hat{b}^*)$ have the same asymptotic variance, so that (6), with b^* replaced by $\hat{b}^*$, is a consistent variance estimator for $\hat{\mu}(\hat{b}^*)$. Indeed it is consistent for the asymptotic variance of any $\hat{\mu}(b)$. In practice, we recommend that one use the estimator $\hat{\mu}(\hat{b}^*)$ in lieu of the alternative estimators $\hat{\mu}(b)$, because then one obtains "for free" a consistent variance estimator. Under regularity conditions, for any given function $l(\cdot)$, $E[l(Y)|\mathbf{V}=\mathbf{v}]$ is consistently estimated by

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

Thus we estimate $b^*(\mathbf{v};\mu)$ by $\hat{b}^*(\mathbf{v};\mu) = \hat{E}[(Y-\mu)\exp(\alpha_0 Y)|\mathbf{V}=\mathbf{v}]/\hat{E}[\exp(\alpha_0 Y)|\mathbf{V}=\mathbf{v}]$. Note that $\hat{\mu}(\hat{b}^*)$ can be written in closed form as

$$\hat{\mu}(\hat{b}^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{\pi}(\mathbf{V}_i,Y_i)}Y_i - \frac{\Delta_i-\hat{\pi}(\mathbf{V}_i,Y_i)}{\hat{\pi}(\mathbf{V}_i,Y_i)}\frac{\hat{E}[Y\exp(\alpha_0 Y)|\mathbf{V}=\mathbf{V}_i]}{\hat{E}[\exp(\alpha_0 Y)|\mathbf{V}=\mathbf{V}_i]}.$$

To summarize the results of this section, we state the following proposition. Here and throughout, results whose proof would require the detailed checking of precise regularity conditions are termed propositions rather than theorems. Rigorous proofs would require modern empirical process theory and are beyond the scope of this article.

Proposition 1. Suppose that $\Lambda_0(T|\mathbf{V}=\mathbf{v})$ is finite, $\mathbf{V}$ is a discrete random vector, and Y has bounded support. Then $\hat{\mu}(\hat{b}^*)$ and $\hat{\mu}(b)$, for any b, are RAL with influence function $-\tau(b^*)^{-1}h(O;\mu_0,\Lambda_0;b^*)$ with asymptotic variance that can be consistently estimated by (6) with b^* replaced by $\hat{b}^*$.

Interestingly, when $\bar{\mathbf{V}}(t)=\mathbf{V}$ is time independent, our estimate $\hat{\Lambda}(T|\mathbf{V}=\mathbf{v})$ of $\Lambda_0(T|\mathbf{V}=\mathbf{v})$ depends on the data only through $\{(\Delta_i, Y_i\Delta_i, \mathbf{V}_i)\colon i=1,\ldots,n\}$. In particular, it is not a function of the actual drop-out times Q or even of their ranks. It follows that for choices of $b(\bar{\mathbf{v}}(t),t;\mu)$ that do not depend on t, including $b^*(\mathbf{v};\mu)$, $\hat{\mu}(b)$ is not a function of the Q_i's. In contrast, it follows from the proof of Theorem 1 in Appendix A that if $\bar{\mathbf{V}}(t)$ were time dependent or if we replaced $\alpha_0 Y$ in (1) by a known function $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)$ that depended on t, then μ_0 would not even be identified in model $A(\alpha_0)$ in the absence of data on the drop-out times Q. Given data on Q, it is straightforward to generalize the estimators $\hat{\mu}(b)$ to obtain estimators of μ_0 in model $A(\alpha_0)$ when $\bar{\mathbf{V}}(t)$ is time dependent (but still low dimensional) and/or $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)$ replaces $\alpha_0 Y$. In this setting the function $b^*(\bar{\mathbf{v}}(t),t;\mu)$, for which $h(O;\mu_0,\Lambda_0;b^*)$ is uncorrelated with all nuisance scores, will depend on t and is characterized in Appendix B.

When $\bar{\mathbf{V}}(t)$ is low dimensional, an obvious competitor to our AIPCW estimator $\hat{\mu}(\hat{b}^*)$ is the nonparametric maximum likelihood estimator (NPMLE) of μ_0 (van der Laan 1993), which is asymptotically equivalent to $\hat{\mu}(\hat{b}^*)$. Indeed, it may be algebraically equivalent, depending on which of several possible "nonparametric likelihood functions" is maximized (Murphy 1995). However, as shown in Section 4, the AIPCW methodology generalizes straightforwardly to model $B(\alpha_0)$ with $\bar{\mathbf{V}}(t)$ high dimensional. In this latter setting, the NPMLE is undefined (Robins and Ritov 1997).

4. ESTIMATION IN MODEL $B(\alpha_o)$

4.1 A Class of Estimators

To motivate our estimator of $\boldsymbol{\psi}_0 = (\mu_0,\boldsymbol{\gamma}_0')'$ in model $B(\alpha_0)$, suppose for the moment that the baseline hazard $\lambda_0(t)$ in (2) is known. Let $\Lambda_0(t)=\int_0^t \lambda_0(u)\,dt$ denote the cumulate baseline hazard. We assume that $\mathbf{W}(t)$ and $\boldsymbol{\gamma}_0$ in (2) are q-dimensional. Consider the $q+1$ vector of augmented IPCW estimating functions $\mathbf{h}(O;\boldsymbol{\psi},\Lambda_0;\mathbf{b})$, where $\mathbf{h}(O;\boldsymbol{\psi},\Lambda_0;\mathbf{b}) = (h_1(O;\boldsymbol{\psi},\Lambda_0,b_1),\ldots,h_{q+1}(O;\boldsymbol{\psi},\Lambda_0,b_{q+1}))'$, $\mathbf{b}=(b_1,\ldots,b_{q+1})$, $b_j=b_j(\bar{\mathbf{v}}(t),t;\boldsymbol{\psi})$ are real-valued functions of $(\bar{\mathbf{v}}(t),t,\boldsymbol{\psi})$ chosen by the data analyst, $h_1(O;\boldsymbol{\psi},\Lambda_0,b_1) = \Delta(Y-\mu)/\pi(\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma}) + a(O;\boldsymbol{\psi},\Lambda_0;b_1)$, for $j\neq 1$, $h_j(O;\boldsymbol{\psi},\Lambda_0,b_j)=a(O;\boldsymbol{\psi},\Lambda_0;b_j)$ with $\pi(\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})\equiv S(T|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})$, $S(t|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})\equiv\exp(-\int_0^t\exp(\boldsymbol{\gamma}'\mathbf{W}(u)+\alpha_0 Y)\lambda_0(u)\,du)$,

$$\begin{aligned} a(O;\boldsymbol{\psi},\Lambda_0;b_j) &\equiv -\frac{\Delta}{\pi(\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})}E_{\boldsymbol{\gamma}}[(1-\Delta)b_j(\bar{\mathbf{V}}(Q),Q;\boldsymbol{\psi})|\bar{\mathbf{V}}(T),Y] \\ &\quad + (1-\Delta)b_j(\bar{\mathbf{V}}(Q),Q;\boldsymbol{\psi}), \end{aligned}$$

and $E_{\boldsymbol{\gamma}}[\cdot|\bar{\mathbf{V}}(T),Y]$ indicates expectations with respect to the distribution $F(t|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma}) = 1-S(t|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})$. Because $E[a(O;\boldsymbol{\psi}_0,\Lambda_0;b_j)]=0$, the estimating function $\mathbf{h}(O;\boldsymbol{\psi}_0,\Lambda_0,\mathbf{b})$ has mean 0. Thus, by standard Taylor series arguments, the solution $\tilde{\boldsymbol{\psi}}(b)$ to

$$\sum_{i=1}^{n} \mathbf{h}(O_i; \boldsymbol{\psi}, \Lambda_0; \mathbf{b}) = 0$$

will be asymptotically normal with mean 0 and asymptotic variance $\boldsymbol{\tau}(\mathbf{b})^{-1}\text{var}[\mathbf{h}(O; \boldsymbol{\psi}_0, \Lambda_0; \mathbf{b})]\boldsymbol{\tau}(\mathbf{b})^{-1'}$ provided that $\boldsymbol{\tau}(\mathbf{b}) \equiv \partial E[\mathbf{h}(O; \boldsymbol{\psi}, \Lambda_0; \mathbf{b})]/\partial \boldsymbol{\psi}_{|\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ is invertible. The asymptotic variance can be consistently estimated by

$$\tilde{\boldsymbol{\tau}}(\mathbf{b})^{-1} \left[n^{-1} \sum_{i=1}^{n} \mathbf{h}(O_i; \tilde{\boldsymbol{\psi}}(b), \Lambda_0; \mathbf{b})^{\otimes 2} \right] \tilde{\boldsymbol{\tau}}(\mathbf{b})^{-1'}, \qquad (8)$$

with $\tilde{\boldsymbol{\tau}}(\mathbf{b}) = n^{-1} \sum_{i=1}^{n} \partial \mathbf{h}(O_i; \tilde{\boldsymbol{\psi}}(\mathbf{b}), \Lambda_0; \mathbf{b})/\partial \boldsymbol{\psi}$.

Because $\Lambda_0(u)$ is unknown, it must be estimated. If we could always observe Y, then we could replace Λ_0 in the foregoing estimating functions by Breslow's profile estimator for the baseline hazard with censoring times representing the jump times (Andersen et al. 1993). We can express this profile estimator as

$$\tilde{\Lambda}(t; \boldsymbol{\gamma}) = \int_0^t \left(\frac{1}{n} \sum_{i=1}^{n} \exp(\boldsymbol{\gamma}'\mathbf{W}_i(u) + \alpha_0 Y_i) I(Q_i \geq u) \right)^{-1} \times \left(\frac{1}{n} \sum_{i=1}^{n} dN_i(u) \right),$$

where $N_i(u) = I(Q_i \leq u, \Delta_i = 0)$. Because Y is not always observed, we modify the foregoing estimator using a similar argument as in Section 3 to yield the following recursive profile estimator for $\Lambda_0(t)$:

$$\hat{\Lambda}(t; \boldsymbol{\gamma}) = \int_0^t \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i \exp(\boldsymbol{\gamma}'\mathbf{W}_i(u) + \alpha_0 Y_i) I(Q_i \geq u)}{\exp(-\int_u^T \exp(\boldsymbol{\gamma}'\mathbf{W}_i(x) + \alpha_0 Y_i)\, d\hat{\Lambda}(x; \boldsymbol{\gamma}))} \right)^{-1} \times \left(\frac{1}{n} \sum_{i=1}^{n} dN_i(u) \right).$$

We can obtain an explicit solution for $\hat{\Lambda}(t; \boldsymbol{\gamma})$ using an approach analogous to that described in Section 3. Given $\hat{\Lambda}(t; \boldsymbol{\gamma})$, we can estimate $S(t|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma})$ by $\hat{S}(t|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma}) = \exp(-\int_0^t \exp(\boldsymbol{\gamma}'\mathbf{W}(u) + \alpha_0 Y)\, d\hat{\Lambda}(u; \boldsymbol{\gamma}))$ and $F(t|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma})$ by $\hat{F}(t|\hat{\mathbf{V}}(T), Y; \boldsymbol{\gamma}) = 1 - \hat{S}(t|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma})$

Now define $\hat{\boldsymbol{\psi}}(\mathbf{b})$ to be the solution to

$$\sum_{i=1}^{n} \mathbf{h}(O_i; \boldsymbol{\psi}, \hat{\Lambda}(\cdot; \boldsymbol{\gamma}); \mathbf{b}) = 0,$$

where $\mathbf{h}(O; \boldsymbol{\psi}, \hat{\Lambda}(\cdot; \boldsymbol{\gamma}); \mathbf{b})$ is defined like $\mathbf{h}(O; \boldsymbol{\psi}, \Lambda_0; \mathbf{b})$ except that S and F are replaced by $\hat{S}$ and $\hat{F}$, so $\hat{E}_{\boldsymbol{\gamma}}[(1 - \Delta) b_j(\bar{\mathbf{V}}(Q), Q; \boldsymbol{\psi})|\bar{\mathbf{V}}(T), Y] = \int_0^T b_j(\bar{\mathbf{V}}(t), t; \boldsymbol{\psi})\, d\hat{F}(t|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma})$ and $\hat{\pi}(\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma}) = \hat{S}(T|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma})$. In the foregoing estimating equation, the function $\mathbf{b}$ need be evaluated only at the censoring times. One would expect that under mild regularity conditions, $\hat{\boldsymbol{\psi}}(\mathbf{b})$ would be a regular asymptotically linear estimator of ψ_0 with influence function $\mathbf{d}(O; \mathbf{b})$, say. However, the asymptotic variance $E[\mathbf{d}(O; \mathbf{b})^{\otimes 2}]$ of $\hat{\psi}(\mathbf{b})$ will not in general be given by an analog of (8), because we need to adjust for the estimation of $\Lambda_0(t)$. Hence further work will be required to obtain confidence intervals for ψ_0. We consider three variance estimation procedures.

The first procedure is to follow the approach of Section 3 and determine those functions $\mathbf{b}^*(\bar{\mathbf{v}}(t), t; \boldsymbol{\psi})$ for which $\mathbf{h}(O; \boldsymbol{\psi}_0, \Lambda_0; \mathbf{b}^*)$ is uncorrelated with the scores S_η for all parametric models $\lambda(t; \eta)$ for $\lambda_0(t)$ of (2), so that no adjustment to the variance will be required. As in the previous case, $\mathbf{b}^*$ will have to be estimated from the observed data. If $\hat{\mathbf{b}}(\bar{\mathbf{v}}(t), t; \boldsymbol{\psi})$ converges in probability to $\mathbf{b}^*(\bar{\mathbf{v}}(t), t; \boldsymbol{\psi})$, then, under mild regularity conditions, the estimator $\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*)$ will be a RAL estimator with asymptotic variance that can be consistently estimated by the following analog of (8):

$$\hat{\boldsymbol{\tau}}(\hat{\mathbf{b}}^*)^{-1} \times \left[n^{-1} \sum_{i=1}^{n} \mathbf{h}(O_i; \hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*), \hat{\Lambda}(\cdot, \hat{\boldsymbol{\gamma}}(\hat{\mathbf{b}}^*)); \hat{\mathbf{b}}^*)^{\otimes 2} \right] \hat{\boldsymbol{\tau}}(\hat{\mathbf{b}}^*)^{-1}, \qquad (9)$$

where $\hat{\boldsymbol{\tau}}(\hat{\mathbf{b}}^*) = n^{-1} \sum_{i=1}^{n} \partial \mathbf{h}(O_i; \hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*), \hat{\Lambda}(\cdot, \hat{\boldsymbol{\gamma}}(\hat{\mathbf{b}}^*)); \hat{\mathbf{b}}^*)/\partial \boldsymbol{\psi}$. The second approach is to develop an analytic expression for the influence function $\mathbf{d}(O; \mathbf{b})$ of $\hat{\boldsymbol{\psi}}(\mathbf{b})$ for any choice of $\mathbf{b}$. In model $B(\alpha_0)$ this approach is somewhat complex, and it is not considered further in this article. The third approach is to recognize that if $\hat{\boldsymbol{\psi}}(\mathbf{b})$ is a RAL estimator, then we can obtain a consistent estimate of its asymptotic variance by the nonparametric bootstrap (Gill 1989). Because in conducting a sensitivity analysis it is necessary to calculate confidence intervals for μ_0 for many values of the selection bias parameter α_0, bootstrap variance estimation may require impractically large computation time. Thus in the simulations and data analyses reported in Sections 5 and 6, we use the first of the three approaches, which we describe in the next subsection. However, it should be noted that the bootstrap variance estimator, in contrast to the analytic estimator (9), will remain a consistent estimator of the asymptotic variance even under misspecification of model $B(\alpha_0)$.

4.2 Estimation of $\mathbf{b}^*$

In model $B(\alpha_0)$, in contrast to model $A(\alpha_0)$, the set $\mathbf{b}^*$ of functions $\mathbf{b}^*$ such that $\mathbf{h}(O; \boldsymbol{\psi}_0, \Lambda_0, \mathbf{b}^*)$ is orthogonal to the scores S_η for any parametric model $\lambda(t; \eta)$ has an infinite number of elements. In Appendix B we show that we can map an arbitrary $q+1$-dimensional function $\boldsymbol{\phi} \equiv \boldsymbol{\phi}(\bar{\mathbf{V}}(t), t)$ into a particular member $\mathbf{b}^*$ of the set $\mathbf{b}^*$ by solving the Volterra integral equation,

$$\begin{aligned} \mathbf{b}^*(\bar{\mathbf{V}}(t), t; \boldsymbol{\psi}_0) &= \boldsymbol{\phi}(\bar{\mathbf{V}}(t), t) - E[S(t|\bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma}_0) \\ &\quad \times \exp(\boldsymbol{\gamma}_0'\mathbf{W}(t) + \alpha_0 Y)]^{-1} q_{\mathbf{b}^*, \boldsymbol{\phi}}(t; \boldsymbol{\psi}_0), \qquad (10) \end{aligned}$$

where

$$
\begin{aligned}
&q_{\mathbf{b},\boldsymbol{\phi}}(t;\boldsymbol{\psi}_0)\\
&\quad= E[\boldsymbol{\phi}(\bar{\mathbf{V}}(t),t)S(t|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma}_0)\exp(\boldsymbol{\gamma}_0'\mathbf{W}(t)+\alpha_0 Y)]\\
&\qquad+\left\{\int_0^t E\left[\mathbf{b}(\bar{\mathbf{V}}(u),u;\boldsymbol{\psi})\,dF(u|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma}_0)\right.\right.\\
&\qquad\qquad\left.\left.\times\ \exp(\boldsymbol{\gamma}_0'\mathbf{W}(t)+\alpha_0 Y)\right]\right\}\\
&\qquad- E[e_1(Y-\mu_0)\exp(\boldsymbol{\gamma}_0'\mathbf{W}(t)+\alpha_0 Y)]
\end{aligned}
$$

and $\mathbf{e}_1$ is the $q+1$-dimensional vector whose first component is 1 and whose remaining components are 0. We also show that any $\mathbf{b}^* \in \mathbf{b}^*$ satisfies (10) for some $\boldsymbol{\phi}(\bar{\mathbf{v}}(t),t)$.

As in Section 3, because the solution $\mathbf{b}^*$ to (10) depends on the unknown distribution F_O of the data, $\mathbf{b}^*$ will have to be estimated. In practice, one selects a function $\boldsymbol{\phi}(\bar{\mathbf{V}}(t),t)$, then obtains an estimator $\hat{\mathbf{b}}^*(\bar{\mathbf{V}}(t),t;\boldsymbol{\psi})$ for the corresponding function $\mathbf{b}^*(\bar{\mathbf{V}}(t),t;\boldsymbol{\psi})$ that is consistent at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$. Let $Q_{(j)}$ be the jth ordered censoring time and define $Q_{(0)}=0$. Then the estimator $\hat{\mathbf{b}}^*(\bar{\mathbf{V}}(t),t;\boldsymbol{\psi})$ is recursively defined (in forward time) by the following empirical version of (10). For $t\in(Q_{(k)},Q_{(k+1)}], k=0,1,2,\ldots,$

$$
\begin{aligned}
&\hat{\mathbf{b}}^*(\bar{\mathbf{V}}(t),t;\boldsymbol{\psi})\\
&\quad=\boldsymbol{\phi}(\bar{\mathbf{V}}(t),t)-\hat{E}_\gamma[\hat{S}(t|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})\\
&\qquad\times\ \exp(\boldsymbol{\gamma}'\mathbf{W}(t)+\alpha_0 Y)]^{-1}\hat{q}_{\hat{\mathbf{b}}^*,\boldsymbol{\phi}}(t;\boldsymbol{\psi}),
\end{aligned}
\tag{11}
$$

where, for any $Z=z(\bar{\mathbf{V}}(T),Y), \hat{E}_\gamma(Z)\equiv n^{-1}\sum_{i=1}^n \Delta_i$ $Z_i/\hat{\pi}(\bar{\mathbf{V}}_i(T),Y_i;\boldsymbol{\gamma}); \hat{\pi}, \hat{F}$, and $\hat{S}$ are as defined earlier;

$$
\begin{aligned}
&\hat{q}_{\mathbf{b},\boldsymbol{\phi}}(t;\boldsymbol{\psi})\\
&\quad=\hat{E}_\gamma[\boldsymbol{\phi}(\bar{\mathbf{V}}(t),t)\hat{S}(t|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})\exp(\boldsymbol{\gamma}'\mathbf{W}(t)+\alpha_0 Y)]\\
&\qquad+\sum_{j=1}^{k}\hat{E}_\gamma[\mathbf{b}(\bar{\mathbf{V}}(Q_{(j)}),Q_{(j)};\boldsymbol{\psi})d\hat{F}(Q_{(j)}|\bar{\mathbf{V}}(T),Y;\boldsymbol{\gamma})\\
&\qquad\qquad\times\ \exp(\boldsymbol{\gamma}'\mathbf{W}(t)+\alpha_0 Y)]\\
&\qquad-\hat{E}_\gamma[\mathbf{e}_1(Y-\mu)\exp(\boldsymbol{\gamma}'\mathbf{W}(t)+\alpha_0 Y)];
\end{aligned}
$$

and $\sum_{j=1}^0\equiv 0$. To execute this recursive algorithm, it is sufficient to have computed $\hat{\mathbf{b}}^*(\bar{\mathbf{V}}(Q_{(j)}),Q_{(j)};\boldsymbol{\psi}), j=0,\ldots,k$ to compute $\hat{\mathbf{b}}^*(\bar{\mathbf{V}}(t),t;\boldsymbol{\psi})$ for all $t\in(Q_{(k)},Q_{(k+1)}]$. In summary, we can state the following proposition.

Proposition 2. Suppose that $\Lambda_0(T)$ is finite, $\mathbf{V}(t)$ is a stochastic process with bounded support, $\boldsymbol{\psi}_0$ lies in the interior of a compact set $\boldsymbol{\Psi}_0\subset R^{q+1}$, and $\hat{\mathbf{b}}^*$ is determined by some function $\boldsymbol{\phi}$ via (11). Then, in model $B(\alpha_0), \hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*)$ is RAL with influence function $-\partial E[\mathbf{h}(O;\boldsymbol{\psi},\Lambda_0;\mathbf{b}^*)]/\partial\boldsymbol{\psi}^{-1}_{|\boldsymbol{\psi}=\boldsymbol{\psi}_0}\mathbf{h}(O;\boldsymbol{\psi}_0,\Lambda_0;\mathbf{b}^*)$ and with asymptotic variance that can be consistently estimated by (9). Here $\mathbf{b}^*$ is the probability limit of $\hat{\mathbf{b}}^*$ at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$.

In Appendix B we show that our class of estimators $\{\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*)\}$ contains, up to asymptotic equivalence, all RAL estimators in model $B(\alpha_0)$. That is, if $\tilde{\boldsymbol{\psi}}$ is any other RAL estimator of $\boldsymbol{\psi}_0$ in model $B(\alpha_0)$, then there will exist some function $\boldsymbol{\phi}=\boldsymbol{\phi}(\bar{\mathbf{v}}(t),t)$ such that $\tilde{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*)$ have the same influence function, with $\hat{\mathbf{b}}^*$ determined by $\boldsymbol{\phi}$ via (11).

4.3 Efficiency

The efficiency of the estimator $\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*)$ will depend on the choice of $\boldsymbol{\phi}$. The optimal choice $\boldsymbol{\phi}_{\text{opt}}$ will result in an estimator $\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*_{\text{opt}})$ whose asymptotic variance will attain the semiparametric variance bound for model $B(\alpha_0)$. Furthermore, because model $B(\alpha_0)$ does not suffer from the curse of dimensionality due to the dimension reduction implicit in (2), the finite-sample variance of $\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*_{\text{opt}})$ should be generally close to the variance predicted by asymptotic theory (Robins and Ritov 1997). Unfortunately, the optimal choice $\boldsymbol{\phi}_{\text{opt}}$ is not available for two reasons. First, $\boldsymbol{\phi}_{\text{opt}}$ is a function of the unknown distribution joint distribution F_O of the observed data. Second, even if F_O were known, calculation of $\boldsymbol{\phi}_{\text{opt}}$ would require solving an exceedingly complex integral equation. In light of this second reason, we forego trying to obtain a semiparametric efficient estimator. A simple choice of $\boldsymbol{\phi}$ that we use in our simulation studies and in our reanalysis of the ACTG 175 data in Sections 5 and 6 is to take

$$
\boldsymbol{\phi}(\bar{\mathbf{V}}(t),t;\boldsymbol{\psi})=(0,\mathbf{W}(t))'. \tag{12}
$$

Although not efficient, this choice of $\boldsymbol{\phi}$ will be suitable when the uncertainty in the mean μ_0 due to not knowing the true value of α_0 is considered by subject matter experts to dominate the uncertainty due to sampling variability. However, in settings where sampling variability dominates, it will often be useful to attempt to find more efficient choices for $\boldsymbol{\phi}$. To this end, in Appendix C we propose an adaptive choice for $\boldsymbol{\phi}, \hat{\boldsymbol{\phi}}_{\text{adap}}$, that will result in highly, although not fully, efficient estimates of the mean μ_0 of Y under model $B(\alpha_0)$. Our approach is motivated by the observations that (a) if $\alpha_0=0$, the semiparametric variance bounds in models $A(\alpha_0)$ and $B(\alpha_0)$ will be identical (Robins and Rotnitzky 1992), and (b) even when $\alpha_0\neq 0$, if $\mathbf{W}(t)$ in (2) is high dimensional, the semiparametric variance bound in model $B(\alpha_0)$ will be only slightly less than the variance bound for the larger model $A(\alpha_0)$. Thus if we can obtain an estimator of μ_0 whose asymptotic variance is close to the variance bound for model $A(\alpha_0)$, then it should have reasonably good efficiency relative to the semiparametric efficient estimator $\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*_{\text{opt}})$ for model $B(\alpha_0)$.

5. SIMULATION STUDIES

To evaluate the finite-sample performance of our estimation techniques under models $A(\alpha_0)$ and $B(\alpha_0)$, we conducted two simulation studies.

5.1 Model $\boldsymbol{A(\alpha_o)}$

We generated data under the assumption that $\mathbf{V}$ was a Bernoulli random variable with mean .3 and that the conditional law of Y given $\mathbf{V}$ was normally distributed with mean $\mathbf{V}-.3$ and variance 1, truncated at $\mathbf{V}-2.26$ and $\mathbf{V}+1.66$. These assumptions imply that the marginal mean of Y is 0. We also assumed that the conditional law of Q

given $\mathbf{V}$ and Y follows an exponential distribution with hazard $(\delta_0 + \delta_1\mathbf{V})\exp(\alpha_0 Y)$, where δ_0 and δ_1 are fixed constants. We took $T = 1$. We selected δ_0, δ_1, and α_0 so that $P[Q \geq 1|\mathbf{V} = 0, Y = -.3] = .65, P[Q \geq 1|\mathbf{V} = 1, Y = .7] = .50$, and $P[Q \geq 1|\mathbf{V} = 1, Y = 2.35] = .40$. These constraints yield true values of δ_0, δ_1, and α_0 of .4308, .1849, and .1691, indicating that subjects with high values of Y and $\mathbf{V}$ were more likely to drop out. We simulated 500 datasets of 500 subjects each. To reflect the fact that in practice one would not know the true value of α_0, we fit five models with equally spaced α_0's ranging from $-.1691$ to $.5073$. The results of this simulation study are presented in Table 1. For each α_0, the table displays the averages of the parameter estimates, the standard deviation of the parameter estimates, and the averages of the standard errors. We see that when we guess the true α_0, we get an unbiased estimate of the mean. Note that our variance estimator appears virtually unbiased for all values of α_0. This reflects the fact that model $A(\alpha_0)$ is, as indicated by Theorem 1, a nonparametric model for the distribution F_O of the observed data. This implies that our variance estimate will be consistent for all values of α_0, not just for the value that generated the data.

5.2 Model $\boldsymbol{B(\alpha_o)}$

For this simulation study, we conceived of a longitudinal study in which measurements were taken at five time points, $t = 0, .25, .5, .75, 1$. Let $\mathbf{V}_t$ denote the measurement at time t. We are interested in making inference about the mean of the measurement $Y = \mathbf{V}_1$ at time $T = 1$. We generated data under the assumption that the measurements were multivariate normal with mean 0, variance 1, and an AR-1 covariance structure in which the covariance between $\mathbf{V}_s$ and $\mathbf{V}_t$ was equal to $.6^{4|s-t|}$. We truncated the measurements at -1.96 and 1.96. So the true mean of Y is 0. We assumed that the measurements were constant between measurements times so that

$$\mathbf{V}(t) = \mathbf{V}_0 I(0 \leq t < .25) + \mathbf{V}_{.25} I(.25 \leq t < .5) \\ + \mathbf{V}_{.5} I(.5 \leq t < .75) + \mathbf{V}_{.75} I(.75 \leq t < 1) + \mathbf{V}_1 I(t \geq 1).$$

In model $B(\alpha_0)$ we chose $\lambda_Q(t|\bar{\mathbf{V}}(1), Y) = \lambda_0(t)\exp(\boldsymbol{\gamma}_0\mathbf{V}(t) + \alpha_0 Y)$; that is, $\mathbf{w}(t, \bar{\mathbf{V}}(t)) = \mathbf{V}(t)$. In generating Q's, we assumed that the baseline hazard was constant. We selected the baseline hazard, $\boldsymbol{\gamma}_0$, and α_0 so that $\Pr[Q \geq 1|\bar{\mathbf{V}}(1-) = 0, Y = 0] = .65, \Pr[Q \geq 1|\bar{\mathbf{V}}(1-) = 1.645, Y = 0] = .50$, and $\Pr[Q \geq 1|\bar{\mathbf{V}}(1-) = 0, Y = 1.645] = .40$. Thus the true baseline hazard was set equal to .4308, the true $\boldsymbol{\gamma}_0$ equal to .2891, and the true α_0 equal to .4588. We simulated 100 datasets of 500 subjects each. Because in reality we do not know the true value of α_0, we fit five models with equally spaced α_0's ranging from 0 to .9176. In the estimation procedure, we chose the $\boldsymbol{\phi}(\bar{\mathbf{V}}(t), t; \boldsymbol{\psi})$ given by (12). The results of this simulation study are presented in Table 2. For varying levels of α_0, this table displays the averages of the parameter estimates, the standard deviation of the parameter estimates, and the averages of the standard errors. We see that when we guess the true α_0, we get unbiased parameter estimates. Finally, it is encouraging to note that our variance estimator performs relatively well at all values of α_0. Note that because model $B(\alpha_0)$ is not nonparametric, our asymptotic theory only predicts that our variance estimator should perform well at the true value of α_0.

Table 1. Results of Simulation Study for Model A(α_0)

	Fixed α_0				
	−.1691	0	.1691	.3382	.5073
Average	−.1548	−.0791	−.0026	.0747	.1520
Standard deviation	.0584	.0592	.0604	.0618	.0638
Average of standard error	.0565	.0567	.0570	.0574	.0578

Table 2. Results of Simulation Study for Model B(α_0)

		Fixed α_0				
		0	.2294	.4588	.6882	.9176
μ_0	Average	−.1782	−.0926	−.0041	.0869	.1771
	Standard deviation	.0500	.0513	.0539	.0577	.0618
	Average of standard errors	.0514	.0538	.0592	.0685	.0823
$\boldsymbol{\gamma}_0$	Average	.4138	.3524	.2884	.2238	.1622
	Standard deviation	.0887	.0917	.0968	.1047	.1161
	Average of standard errors	.0808	.0811	.0827	.0866	.0934

6. SENSITIVITY ANALYSIS OF ACTG 175

In this section we return to the analysis of ACTG 175 started in Section 2.2. Table 3 presents the estimated means and standard errors for CD4 at week 56 for each of the treatment groups using only the completers; that is, non–drop-outs. We have also included the drop-out rates. One naive way to estimate the mean CD4 count μ_0 at week 56 is to simply take the sample average over the completers. This estimate will be unbiased if the data are missing completely at random (MCAR). Treatment comparisons at week 56 using the naive approach show that AZT is inferior to the other three treatments, with some mild evidence of superiority of AZT + ddI over ddI. We fit models $A(\alpha_0)$ and $B(\alpha_0)$ to these data to see how robust this inference is to violation of the MCAR assumption. Due to space limitations, we provide the results of only two sensitivity analyses, one for model $A(\alpha_0)$ and one for model $B(\alpha_0)$.

6.1 Model $\boldsymbol{A(\alpha_o)}$

As described in Section 2.2, we considered model $A(\alpha_0)$ with $\mathbf{V}(t)$ the time-independent covariate denoting IV drug

Table 3. Comparison of Mean Observed CD4 Counts at Week 56

	CD4 at 56 Weeks		
Treatment	Mean	S.E.	Drop-outs
AZT	312.05	7.11	36.0%
AZT + ddI	384.42	8.54	33.6%
AZT + ddC	369.55	7.71	36.6%
ddI	359.60	7.68	26.5%

user status at baseline. Figure 1 presented the estimated means along with 95% confidence intervals for α_0's ranging from $-.02$ to $.02$ for each of the four treatment groups.

To compare treatment groups, consider Figure 2. Here we present six contour plots, each representing a pairwise treatment comparison. To illustrate the AZT versus AZT + ddI comparison, note that on the x-axis we have varying levels of selection bias for the AZT arm, and on the y-axis we have varying levels of selection bias in the AZT + ddI arm. For each combination of selection biases, we perform a test (at the .05 level) of the null hypothesis of no treatment difference between mean CD4 at week 56. The graph is a contour plot of the Z statistic as a function of the two levels of selection biases. The two lines in each plot represent the combinations that lead to a Z statistic of 1.96 and -1.96. To the left of the -1.96 line, we conclude that the data provide evidence that AZT + ddI is better than AZT, and to the right of the 1.96 line, we conclude that the data favor AZT. Between the lines, there is not enough evidence to draw either conclusion. The point at (0, 0) represents the CAR comparison, which jibes with the MCAR conclusion that AZT + ddI is better than AZT. This plot shows that this conclusion is quite robust. Significant differential selection biases would have to occur to alter this conclusion. For example, we would change our conclusion if the selection bias parameters in the AZT and AZT + ddI arms were .01 and 0.

How did we decide to choose the range of $-.02$– $+.02$ for α_0 in Figure 1? The simple rule, which we followed, is that a sensitivity analysis should include a range of selection bias parameters α_0 that contains all values that would be considered plausible by relevant subject matter experts. To include values of α_0 that lie outside the plausible range does no harm, because subject matter experts will discount the results for values of α_0 outside this range.

6.2 Model $B(\alpha_o)$

In model $B(\alpha_0)$ we included CD4 as a time-varying regressor as well as the baseline covariates: age, CD4 count, and IV drug use. As in the simulation study in the previous section, we assume that CD4 counts are constant between measurements. Specifically, we let T =

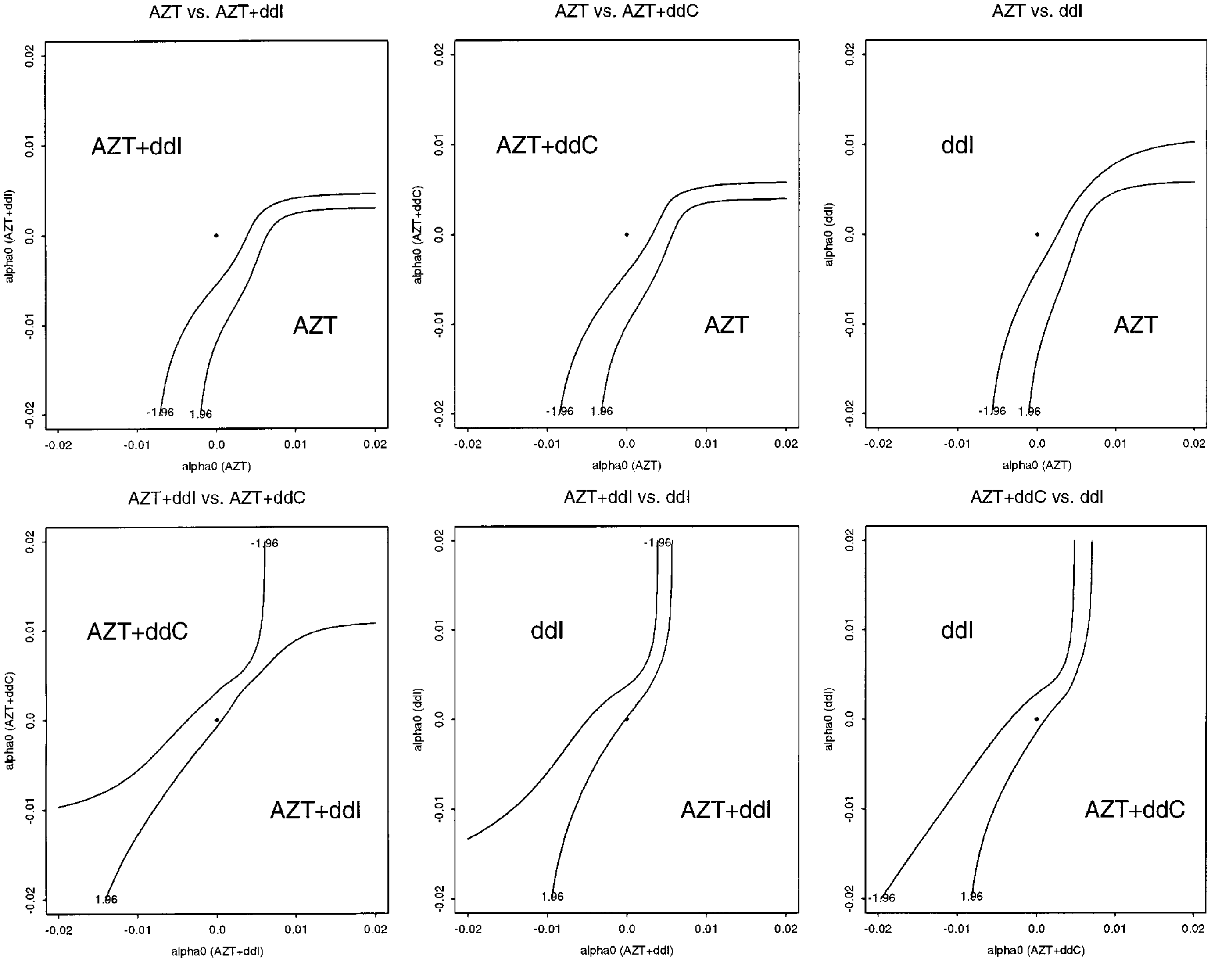

Figure 2. Pairwise Comparison of Treatment Groups in Model A(α_0) with IV Drug User Status as the Time-Independent Regressor.

56, $Z(t)$ denote the CD4 count at time $t, V_0 = Z(0), Y = Z(T), V_1$ denote age, V_2 denote IV drug user status, and $\mathbf{V}(t) = (V_0, V_1, V_2, Z(t))'$. Then, in model $B(\alpha_0)$ we assume that

$$\lambda_Q(t|\bar{\mathbf{V}}(T), Y) = \lambda_0(u)\exp(\gamma_{00}V_0 + \gamma_{01}V_1 + \gamma_{02}V_2 + \gamma_{03}Z(t) + \alpha_0 Y).$$

In the estimation procedure, we chose $\phi(\bar{\mathbf{V}}(t), t; \psi)$ given by (12). In this setting α_0 is interpreted as the log hazard ratio of nonresponse between patients who have the same covariate history, but differ by 1 CD4 count at week 56. When $\alpha_0 = 0$, the Y's are CAR. When $\alpha_0 > 0$, we are assuming that among subjects with the same covariate history, those with higher values of Y are more likely to drop out. The opposite interpretation holds when $\alpha_0 < 0$. Figures 3 and 4 are the exact analogs of Figures 1 and 2 for this model. In general, we include time-independent and time-dependent covariates in $\bar{\mathbf{V}}(t)$ that are correlated with the outcome Y and may predict drop-out at t in the hopes of making the selection process approximately ignorable. (See Sec. 7.2 for further discussion of this matter.)

The conclusions based on Figures 3 and 4 are qualitatively the same as those based on Figures 1 and 2. That is, significant differential selection bias would have to occur for us to change our inference about AZT relative to the other three treatments, and inference about the other treatment comparisons is highly sensitive to nonignorability. However, the sensitivity of the estimated mean to comparable changes in α_0 is rather less in Figures 3 and 4 than in Figures 1 and 2. Specifically, the variation in the estimated mean CD4 count at week 56 as α_0 varies from $-.02$ to $.02$ is less in Figure 3 than in Figure 1. Similarly, when we restrict α_0 in the arms being compared to the interval $(-.005, .005)$, we observe that in Figure 2 but not in Figure 4 there are small regions where AZT is preferred to the other treatments. In Section 7.2 we consider possible explanations for these observations.

7. ADDITIONAL CONSIDERATIONS

In this section we take up a number of remaining issues, several of which were raised by the referees.

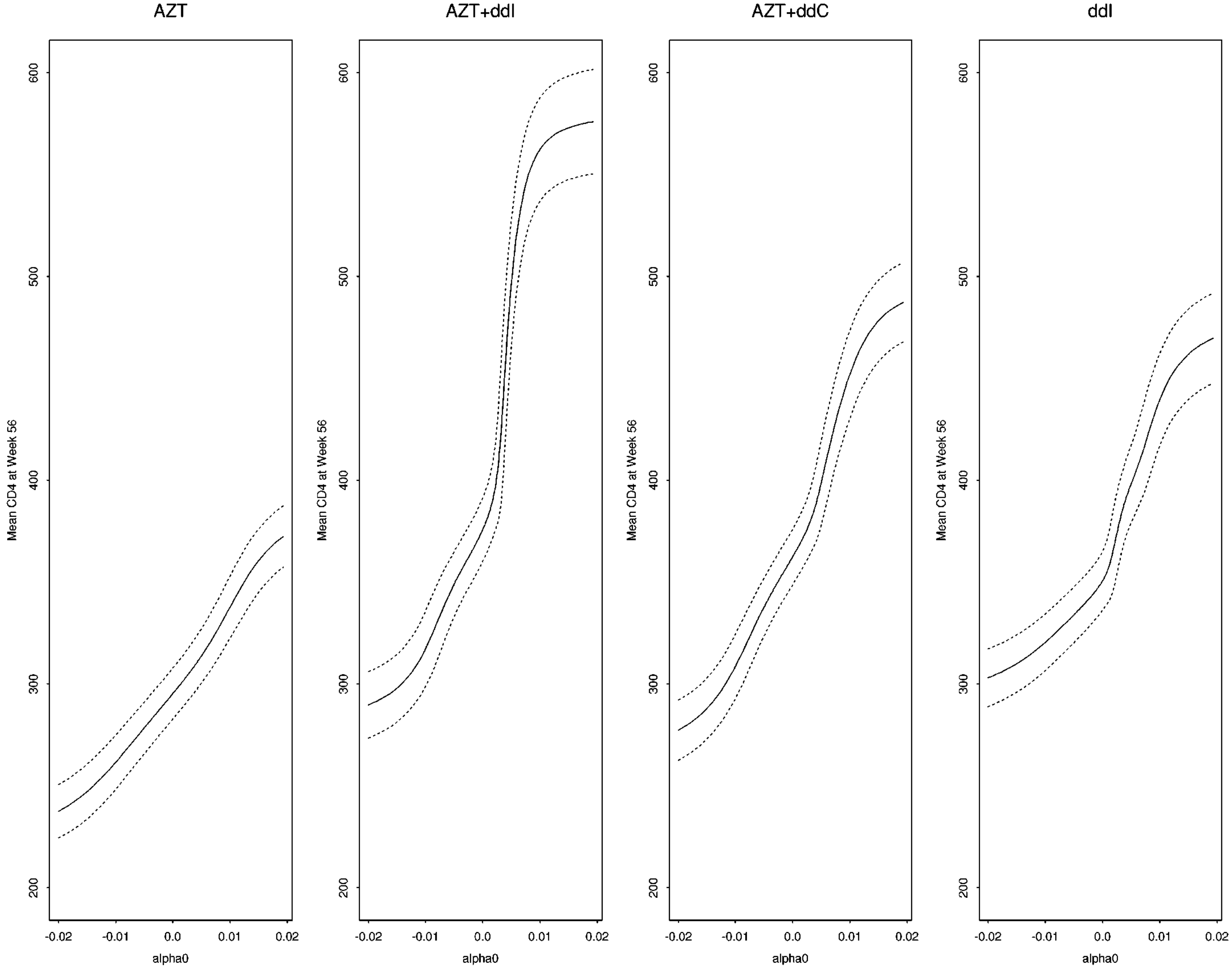

Figure 3. Treatment-Specific Predicted Means and 95% Confidence Intervals for Varying α_0's in Model $B(\alpha_0)$ With Baseline CD4, Age, IV Drug User Status, and Time-Dependent CD4 as the Regressors.

7.1 Estimation of Other Smooth Functionals of the Distribution of $\boldsymbol{Y}$

It is easy to adapt the estimation procedures described in Sections 3 and 4 to estimate other smooth functionals of the marginal distribution of Y. One simply needs to replace $(Y - \mu)$ in the equations of these sections with the full data influence function for the functional of interest. For example, if we would like to estimate the median of Y, then we replace $(Y - \mu)$ by $I(Y \geq \mu) - .5$, where μ now denotes the median. For influence functions that are not differentiable in μ, we suggest using numerical derivatives to estimate $\tau(b)$ in the asymptotic variance. In model $A(\alpha_0)$ we estimated the median as a function of α_0 for each of the four treatment groups. As expected, for each treatment group, the estimated medians increased monotonically with α_0. For positive values of α_0, the rate of increase was less for the median than for the mean. This observation can be explained as follows. Because the empirical distributions of the observed Y's have short left tails but long right tails, we would expect that the mean, but not the median, would be highly sensitive to the assumption, encoded in a large positive value of α_0, that the drop-outs consist largely of subjects with values of Y in the extreme right tail.

7.2 Bounds and the Breakdown of Augmented Inverse Probability of Censoring Weighted Estimators

7.2.1 Bounds. An alternative to our approach based on sensitivity analysis is one based on estimating upper and lower bounds for the mean μ_0 compatible with the observed data. Specifically, if Y is a bounded random variable and the upper and lower bounds are known, then one obtains an estimate $\hat{\mu}_{\text{upper}}$ $(\hat{\mu}_{\text{lower}})$ of the upper (lower) bound for μ_0 by filling in the unobserved Y's with the largest (smallest) possible value of variable $Y, y_{\max}$ $(y_{\min})$. It seems natural to hope that our choice of selection bias function r $(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ attains these bounds in the sense that, as α_0 approaches infinity and minus-infinity, our estimates of μ_0 approach the estimated upper and lower bounds just described. In model $A(\alpha_0)$ our choice of $\alpha_0 Y$ for $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ often satisfies this hope even in finite samples when the random variable $\bar{\mathbf{V}}(T)$ is discrete with only a moderate number of levels. Specifically, suppose that $\mathbf{V}(t)$

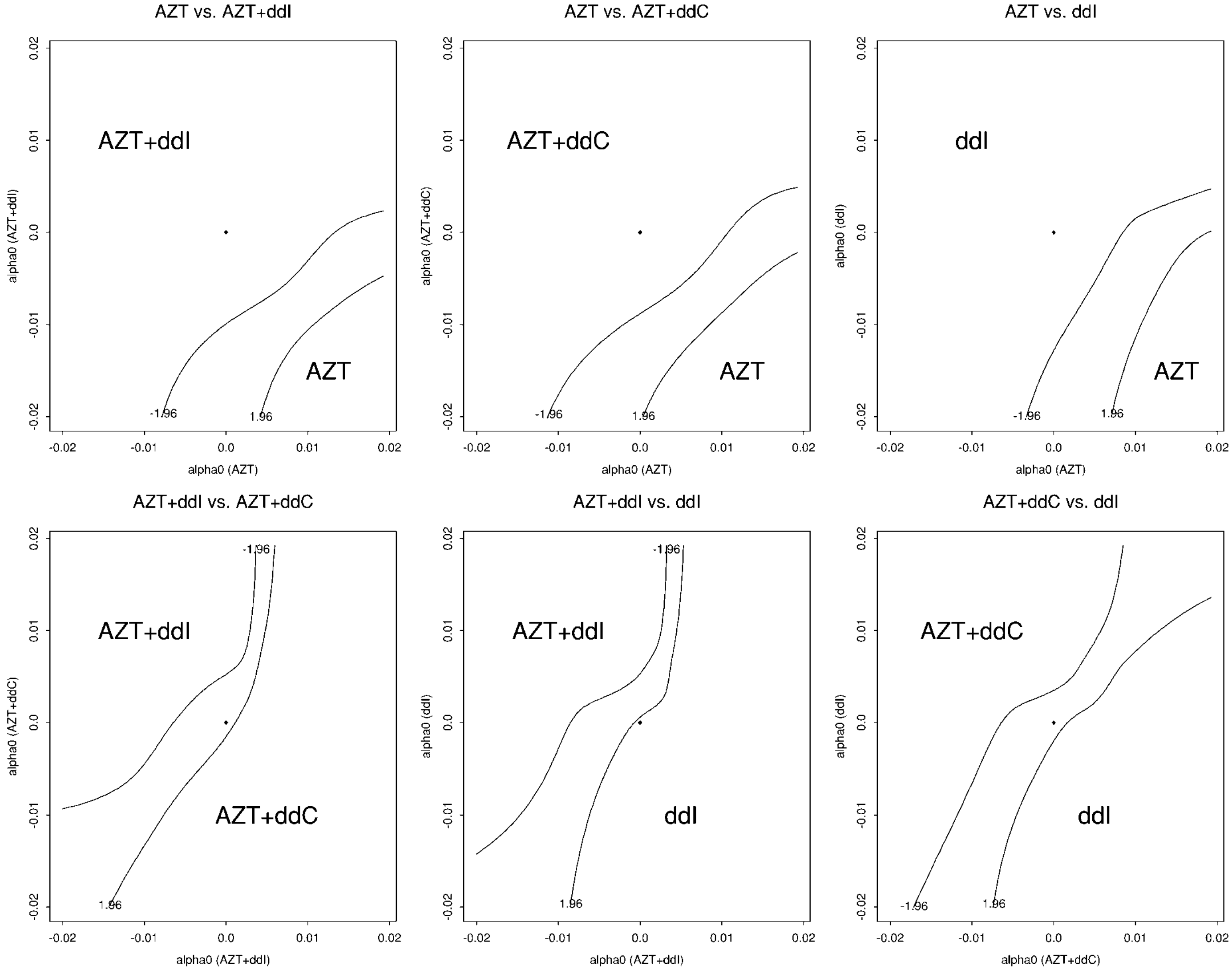

Figure 4. Pairwise Comparison of Treatment Groups in Model $B(\alpha_0)$ with Baseline CD4, Age, IV Drug User Status, and Time-Dependent CD4 as the Regressors.

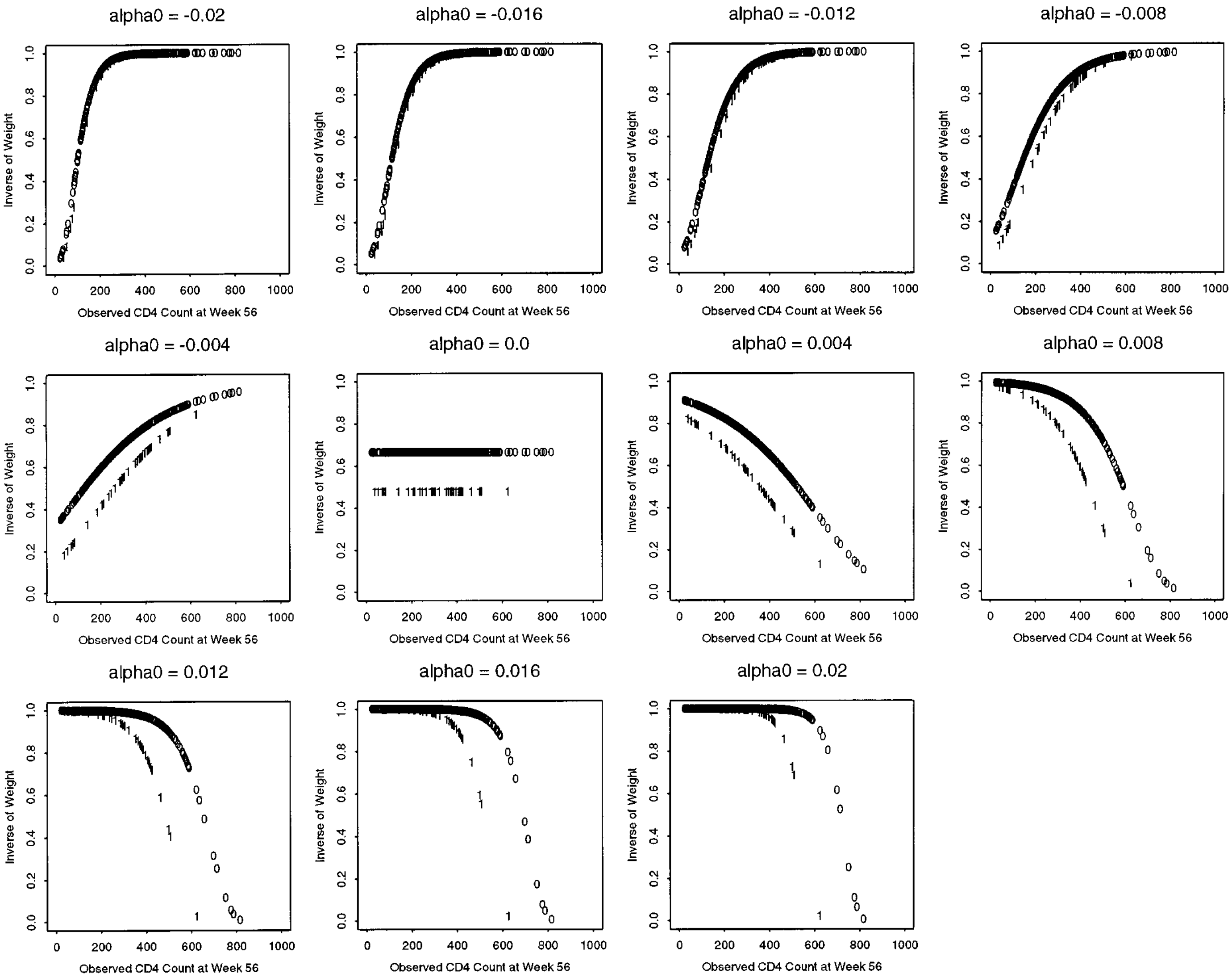

Figure 5. Observed CD4 at Week 56 Versus the Inverse of the Associated Estimated Weights in Model A(α_0), Stratified by IV Drug User Status (1, IV Drug User; 0, Non—Drug User).

is discrete for each t and the number of potential jump times for the process $\bar{\mathbf{V}}(t)$ is small. Then, if at each possible level $\bar{\mathbf{v}}(T)$ of $\bar{\mathbf{V}}(T)$ there is a subject in the dataset whose observed value of Y attains the upper (lower) bound for Y, then the estimates $\hat{\mu}(b)$ and $\hat{\mu}(\hat{b}^*)$ of Section 3 will approach $\hat{\mu}_{\text{upper}}$ ($\hat{\mu}_{\text{lower}}$) as $\alpha_0 \to \infty$ ($\alpha_0 \to -\infty$).

To see why, we study the simplest case. Specifically, we consider the behavior of the IPCW estimator $\hat{\mu}(0)$ solving the estimating equation $0 = \sum_i \Delta_i(Y_i - \mu)/\hat{\pi}(\mathbf{V}_i, Y_i)$ of Section 3, when $\bar{\mathbf{V}}(t) = \mathbf{V}$ is time independent and discrete and $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y) = \alpha_0 Y$. Each completer (non–drop-out) contributes a weight $\hat{\pi}^{-1}$ that depends on α_0. When $\alpha_0 = 0$, each completer in stratum $\mathbf{V} = \mathbf{v}$ receives weight $\hat{\pi}^{-1} = 1 + n^{\mathbf{v}}_{\text{drop}}/n^{\mathbf{v}}_{\text{complete}}$, where $n^{\mathbf{v}}_{\text{drop}}$ and $n^{\mathbf{v}}_{\text{complete}}$ are the number of drop-outs and the number of completers in stratum $\mathbf{v}$. This is because in stratum $\mathbf{v}$ we need to redistribute the contribution of the $n^{\mathbf{v}}_{\text{drop}}$ drop-outs to the completers. When $\alpha_0 = 0$, within stratum $\mathbf{v}$ all completers are exchangeable. So we redistribute the drop-outs' contribution equally among the $n^{\mathbf{v}}_{\text{complete}}$ completers. Thus each completer receives a weight 1 (corresponding to themselves) and an additional weight $n^{\mathbf{v}}_{\text{drop}}/n^{\mathbf{v}}_{\text{complete}}$ to account for the drop-outs. Because model $A(\alpha_0)$ leaves the baseline hazard $\lambda_0(t|\mathbf{V})$ unrestricted, all redistribution of weight is stratum specific. For $\alpha_0 \neq 0$, in stratum $\mathbf{v}$ the fraction of the drop-outs' total contribution $n^{\mathbf{v}}_{\text{drop}}$ assigned to a completer will depend on the completer's outcome Y. Let $y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$) be the maximum (minimum) of the observed values of Y among completers in stratum $\mathbf{v}$, and let $n^{\mathbf{v}}_{\max}$ ($n^{\mathbf{v}}_{\min}$) be the number of completers in stratum $\mathbf{v}$ with $Y = y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$). Then as $\alpha_0 \to \infty$ ($\alpha_0 \to -\infty$), completers in stratum $\mathbf{v}$ with $Y = y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$) are assigned weights tending to $\hat{\pi}^{-1} = 1 + n^{\mathbf{v}}_{\text{drop}}/n^{\mathbf{v}}_{\max}$ ($\hat{\pi}^{-1} = 1 + n^{\mathbf{v}}_{\text{drop}}/n^{\mathbf{v}}_{\min}$); completers whose observed Y do not equal $y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$) are assigned weights tending to 1. The intuition is as follows. Consider two values y_1 and y_2 of Y with $y_2 > y_1$. Then within stratum $\mathbf{v}$, the hazard ratio for drop-out of a subject with $Y = y_2$ compared to a subject with $Y = y_1$ is $\exp(\alpha_0(y_2 - y_1))$, which goes to $\infty(0)$ as $\alpha_0 \to \infty$ ($\alpha_0 \to -\infty$). Thus when $\alpha_0 \to \infty$ ($\alpha_0 \to -\infty$), our estimation assigns any drop-out the largest (smallest) possible value of Y in the stratum; that is, $y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$). It follows that as $\alpha_0 \to \infty$ ($\alpha_0 \to -\infty$), $\hat{\mu}(0)$

[and indeed $\hat{\mu}(\hat{b}^*)$ and $\hat{\mu}(b)$ for any b] tends to $\hat{\mu}_{\max}$ ($\hat{\mu}_{\min}$), where $\hat{\mu}_{\max}$ ($\hat{\mu}_{\min}$) is the sample average of Y over the n study subjects when the drop-outs in stratum $\mathbf{v}$ have been imputed the common value $y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$). If $y^{\mathbf{v}}_{\max} = y_{\max}$ ($y^{\mathbf{v}}_{\min} = y_{\min}$) for all $\mathbf{v}$, then $\hat{\mu}_{\max}$ and $\hat{\mu}_{\min}$ will equal the upper and lower bounds $\hat{\mu}_{\text{upper}}$ and $\hat{\mu}_{\text{lower}}$.

To illustrate the foregoing discussion of weights, we return to our analysis of the ACTG 175 data under model $A(\alpha_0)$. Figure 5 plots, for the AZT treatment arm, the inverse weights $\hat{\pi}$ for the completers as a function of the observed outcome Y and IV drug user status $\mathbf{V}$ for different values of α_0. As expected, when α_0 is very positive (negative), the weights $\hat{\pi}^{-1}$ greatly exceed 1 for only the few subjects with extremely large (small) values for Y. Further, as seen in Figure 1, as α_0 becomes very positive (negative), $\hat{\mu}(\hat{b}^*)$ approaches an asymptote equal to $\hat{\mu}_{\max}$ ($\hat{\mu}_{\min}$). The weights in our simulation experiment did not blow up similarly, because of our truncation of the range of Y.

7.2.2 Breakdown of Augmented Inverse Probability of Censoring Weighted Estimators. It is well known that the performance of IPCW and AIPCW estimators such as $\hat{\mu}(\hat{b}^*)$ can degrade as the weights $\hat{\pi}^{-1}$ become highly skew, because the estimator $\hat{\mu}(\hat{b}^*)$ is then largely determined by those few individuals with large weights. When as in the ACTG 175 data (with $\mathbf{V}$ being IV drug user status), the empirical conditional distributions of Y given $\mathbf{V} = 1$ and $\mathbf{V} = 0$ in the completers are reasonably spread out and have substantial overlap, and α_0 is chosen very positive (negative), the estimated weights $\hat{\pi}^{-1}$ will be markedly skew and highly positively (negatively) correlated with the observed Y. This indicates that it is likely that under the law F_O of the observed data, the population weights $\pi(\mathbf{V}, Y)^{-1} = \Pr[\Delta = 1|\mathbf{V}, Y]^{-1}$ will also be quite skew and highly positively (negatively) correlated with Y given $\Delta = 1$. In such a case the AIPCW estimator of the mean breaks down, because, with high probability, subjects with large (small) values of Y will not be captured in the sample. As a result, with high probability, the estimator $\hat{\mu}(\hat{b}^*)$ will seriously underestimate (overestimate) the mean μ_0, and, furthermore, the variance estimator (6) will severely underestimate the true variability of $\hat{\mu}(\hat{b}^*)$. Indeed, μ_0 is identified under model $A(\alpha_0)$ if and only if $\pi(\mathbf{V}, Y) > 0$ with probability 1, which is equivalent to saying that at each level of $\mathbf{V}$, the support of Y among the drop-outs ($\Delta = 0$) is contained within the support of Y for the completers ($\Delta = 1$). Regularity condition 2 of Appendix A implies that $\pi(\mathbf{V}, Y) > 0$ with probability 1.

One can try to deal with the breakdown of the estimator $\hat{\mu}(\hat{b}^*)$ by a combination of one or more of the following: (a) reassess the substantive plausibility of the values of α_0 causing the trouble; (b) restrict attention to functionals such as the median that are less sensitive to the tails of the distribution of Y; (c) replace $\exp(\alpha_0 Y)$ in (1) by a bounded, less rapidly increasing function of Y; (d) incorporate in the analysis additional time-independent or dependent covariates $\mathbf{V}(t)$ (as in Sec. 6.2); (e) specify a parametric model for the law of $(\mathbf{V}, Y)$ and replace our AIPCW estimator with a parametric likelihood-based estimator (which effectively imputes values of Y to the drop-outs that lie outside the observed sample range); and (f) replace model $A(\alpha_0)$ with an alternative NPI model that naturally allows for extrapolation outside the range of Y in the completers, by not requiring that $\pi(\mathbf{V}, Y) > 0$ with probability 1 for identification of the mean. None of these alternatives is necessarily a satisfactory solution.

For instance, whether strategy (a) is satisfactory will depend on the substantive setting. For example, in ACTG 175, the estimator $\hat{\mu}(\hat{b}^*)$ of Section 6.1 will break down only for values of α_0 that imply that the mean of Y among the drop-outs in at least one stratum $\mathbf{v}$ is nearly as large (small) if not larger (smaller) than $y^{\mathbf{v}}_{\max}$ ($y^{\mathbf{v}}_{\min}$). The AIDS clinicians we have consulted do not find this magnitude of selection bias credible. Suggestion (b) may be unsatisfactory for two reasons. First, the median may not be an estimand of scientific interest. Second, the estimated median might on occasion be surprisingly sensitive. As an extreme but illustrative example, suppose that no covariate data $\mathbf{V}$ are available and that 51% of the subjects drop out. Then as $\alpha_0 \to \infty$ ($\alpha_0 \to -\infty$), the estimate of the median converges to the maximum (minimum) observed Y and will be greater (less) than the estimated mean, indicating greater sensitivity of the median than of the mean to the changes in α_0. Suggestion (c) may be unsatisfactory when, based on subject matter considerations, the exponential form form $\exp(\alpha_0 Y)$ is considered to be more plausible than other forms. However, we used the exponential form in our analysis of ACTG 175, not because we thought it substantively plausible, but rather because it is the usual default choice, and because it can result in the breakdown of AIPCW estimators, opening the door to this very discussion. Suggestion (d) is considered in the next section. Suggestion (e) can be unsatisfactory because, as discussed earlier, assuming a parametric model may result in scientifically unjustified identification of α_0 and μ_0. In fact, we recommend option (f) whenever it is substantively plausible that the support of the distribution of Y among the drop-outs may differ from that among the completers. (See Sec. 7.3.2 for details and caveats.)

7.2.3 Adjustment for Additional Covariates and Plausible Ranges for Sensitivity Analysis Parameters. It is scientifically desirable to adjust for selection bias due to measured covariates by including them in $\bar{\mathbf{V}}(t)$. In this sense, suggestion (d) is always a good one. Because the number and nature of measured factors varies from study to study, it is important that subject matter experts be able to provide a plausible range for α_0 in (1) for various choices of $\bar{\mathbf{V}}(t)$. Adding to $\bar{\mathbf{V}}(t)$ data on additional time-independent and dependent covariates that are both correlated with the outcome Y and predict drop-out at t will usually serve to diminish the degree of nonignorable selection bias due to unmeasured factors. Initially, we had expected this to imply that adding covariates to the analysis would also serve to restrict the range of values of α_0 considered plausible. We were mistaken, because the meaning of the parameter α_0 of the multiplicative hazard model (1) changes when we change the covariates in (1), as we now explain. We use the subscript 1 to indicate models in which the covariate $\bar{\mathbf{V}}(t) = \mathbf{V}$ is the dichotomous covariate IV drug user status and the subscript 2 to denote models in which $\bar{\mathbf{V}}(t)$ is IV drug user status, age, baseline CD4, and current CD4 count. We also use these subscripts to distinguish, when necessary, the selection bias parameter α_0 of model A_1 from that of model A_2 or B_2. Thus α_{02} is the selection bias parameter of models A_2 and B_2. Let $\hat{\mu}_1(\alpha_0)$ and $\hat{\mu}_2(\alpha_0)$ denote the maps from α_0 to $\hat{\mu} \equiv \hat{\mu}(\hat{b}^*)$, depicted in Figures 1 and 3, where in the definition of $\hat{\mu}_1(\alpha_0)$ and $\hat{\mu}_2(\alpha_0)$ we have suppressed their dependence on $\hat{b}^*$. Inspection of these figures reveals that $\hat{\mu}_1(\alpha_0)$ and $\hat{\mu}_2(\alpha_0)$ are both monotone increasing, with $\hat{\mu}_1(\alpha_0)$ increasing more rapidly. Each value of α_{02} determines a unique value of α_{01} through the map $\hat{\mu}_1^{-1}(\hat{\mu}_2(\alpha_{02}))$. This map is evaluated in Table 4 for various values of α_{02}. For the moment, suppose that sampling variability and model misspecification are absent so that models A_2 and B_2 are correct with common parameter α_{02} and model A_1 is correct with parameter α_{01}. Then, by Theorem 1, if α_{02} is the true value of α_0 in (1) that generated the underlying data $(Y, \bar{\mathbf{V}}(T), Q)$ under model A_2 and B_2, then $\alpha_{01} = \mu_1^{-1}(\mu_2(\alpha_{02}))$ must be the true value of α_0 under model A_1, as $\mu_1^{-1}\{\mu_2(\alpha_{02})\}$ is the only value of α_{01} that implies the same mean for Y. It follows that if a subject matter expert has specified a plausible range of, say, $(-.01, .01)$ for α_{02} in the AZT arm, then the expert's plausible range for α_{01}, once the distribution F_O of the data becomes known, is logically fixed at $(\hat{\mu}_1^{-1}(\hat{\mu}_2(-.01)), \hat{\mu}_1^{-1}(\hat{\mu}_2(.01))) = (-.0070, .0025)$. Quite generally if, as in the ACTG 175 data, the slope of $\hat{\mu}_1(\alpha_0)$ is steeper than that of $\hat{\mu}_2(\alpha_0)$, then, as is borne out in Table 4 and contrary to our initial intuition, the length of any plausible range for α_{01} will be narrower than that for α_{02}. Put differently, in the ACTG 175 data, the magnitude of nonignorable selection bias for estimation of μ_0 encoded by $\alpha_{01} = c$ for some nonzero constant c is generally greater than that encoded by $\alpha_{02} = c$. In practice, due to sampling variability and model misspecification or incompatibility, α_{01} will not actually be logically tied to α_{02} through the function $\hat{\mu}_1^{-1}(\hat{\mu}_2(\alpha_{02}))$, but the foregoing discussion should remain qualitatively correct. Incompatibility of models A_1 and A_2 is defined as follows. Given F_O, let $F_1(\alpha_{01})$ and $F_2(\alpha_{02})$ be the laws for $(\mathbf{V}, Y, Q)$ and $(\bar{\mathbf{V}}(T), Y, Q)$ under models $A_1(\alpha_{01})$ and $A_2(\alpha_{02})$, as described in Theorem 1. Let $F_{\text{marginal},2}(\alpha_{02})$ be the marginal law for $(\mathbf{V}, Y, Q)$ induced by $F_2(\alpha_{02})$, with $\mathbf{V}$ being IV drug user status. We then say that model A_1 is incompatible with model A_2 at α_{02} if there exists no value of α_{01} for which $F_1(\alpha_{01})$ equals $F_{\text{marginal},2}(\alpha_{02})$.

Table 4. $\hat{\mu}_1^{-1}(\hat{\mu}_2(\alpha_{02}))$ for Various Values of α_{02}

	$\alpha_{01} = \hat{\mu}_1^{-1}\ (\hat{\mu}_2(\alpha_{02}))$			
α_{02}	*AZT*	*AZT + ddI*	*AZT + ddC*	*ddI*
−.020	−.0110	−.0125	−.0140	−.0103
−.015	−.0090	−.0105	−.0110	−.0083
−.010	−.0070	−.0083	−.0075	−.0060
−.005	−.0040	−.0045	−.0035	−.0040
.000	−.0020	−.0008	−.0008	−.0013
.005	.0000	.0043	.0025	.0033
.010	.0025	.0070	.0048	.0053
.015	.0040	.0080	.0053	.0064
.020	.0045	.0085	.0060	.0068

We now show, somewhat informally, that the ACTG 175 data are not anomalous. Specifically, we argue that when, as in the ACTG 175 data, $\bar{\mathbf{V}}(t)$ and Y are highly correlated among the completers $(\Delta = 1)$ for most times t, we would expect $\hat{\mu}_1(\alpha_0)$ to be steeper than $\hat{\mu}_2(\alpha_0)$. This was first suggested to us by Victor DeGruttola. Informally and qualitatively, we can think of representing the time-independent and time-dependent covariates in model A_2 as a single, discrete covariate $\mathbf{V}$ with many strata. We know that as $\alpha_{02} \to \infty$ $(\alpha_{02} \to -\infty)$, each drop-out in stratum $\mathbf{V} = \mathbf{v}$ will effectively be imputed $y_{\max}^{\mathbf{v}}$ $(y_{\min}^{\mathbf{v}})$, as discussed earlier. If $\mathbf{V}$ is highly correlated with Y among the completers, then the difference in a given stratum $\mathbf{v}$ between $y_{\max}^{\mathbf{v}}$ and $y_{\min}^{\mathbf{v}}$ will be small, and thus we get little difference in the estimated mean for α_{02} very positive versus α_{02} very negative. In contrast, when there is a single dichotomous covariate $\mathbf{V}$, as in model A_1, there will often be a large difference between $y_{\max}^{\mathbf{v}}$ and $y_{\min}^{\mathbf{v}}$, so that, as discussed previously, the estimated mean will depend greatly on whether the drop-outs are assigned $y_{\max}^{\mathbf{v}}$ versus $y_{\min}^{\mathbf{v}}$. Thus, as suggested by Figures 1 and 3, we would expect that for large values of $\alpha_0, \hat{\mu}_1(\alpha_0) - \hat{\mu}_1(-\alpha_0)$ would greatly exceed $\hat{\mu}_2(\alpha_0) - \hat{\mu}_2(-\alpha_0)$ whenever Y is highly correlated with the covariates among the completers $(\Delta = 1)$.

To illustrate the connection between the adjustment for additional covariates and the breakdown of our AIPCW estimators, Figure 6 plots the estimated inverse weights $\hat{\pi}$ as a function of the observed Y's for various values of α_{02} obtained from our fit of model $B_2(\alpha_{02})$ in Section 6.2 to the data for the AZT treatment arm. We note two important differences from Figure 5. For the same value of $\alpha_0 = \alpha_{01} = \alpha_{02}$, both the skewness of the weights and their correlation with the observed Y's are less in model $B_2(\alpha_{02})$ than in model $A_1(\alpha_{01})$, particularly for large positive values of α_0. A full explanation of these differences would require careful consideration of the smoothing effect of model restriction (2) and of the effect of inclusion of time-dependent CD4 count. Here we provide just one possible qualitative explanation for the weight distribution. Our purpose is solely to provide a sense of the issues involved. To this end, we again informally represent $\bar{\mathbf{V}}(t)$ as a single, discrete covariate $\mathbf{V}$ with very many levels. If $\mathbf{V}$ were only weakly predictive of drop-out, then there would be at most a few dropouts at any level. Then, even as $\alpha_0 \to \infty$ $(\alpha_0 \to -\infty)$, the maximal (minimal) weight $1 + n_{\text{drop}}^{\mathbf{v}}/n_{\max}^{\mathbf{v}}$ $(1 + n_{\text{drop}}^{\mathbf{v}}/n_{\min}^{\mathbf{v}})$ assigned to any individual in each stratum $\mathbf{v}$ would not be large. But in Figure 6, at $\alpha_{02} = 0$, we see that $\hat{\pi}$ ranges from about .22 to .85, indicating a rather strong effect of $\mathbf{V}$ on drop-out at time t. Thus the pattern of weights seen in Figure 6 could occur if in addition to $\mathbf{V}$ being highly predictive of drop-out, it was also the case that $y_{\max}^{\mathbf{v}}$ differed markedly across strata of $\mathbf{v}$ but $y_{\min}^{\mathbf{v}}$ did not. Such a distribution would imply that, as seen in Figure 6, when α_0 is very negative, all subjects with large estimated weight have small CD4 counts, but when

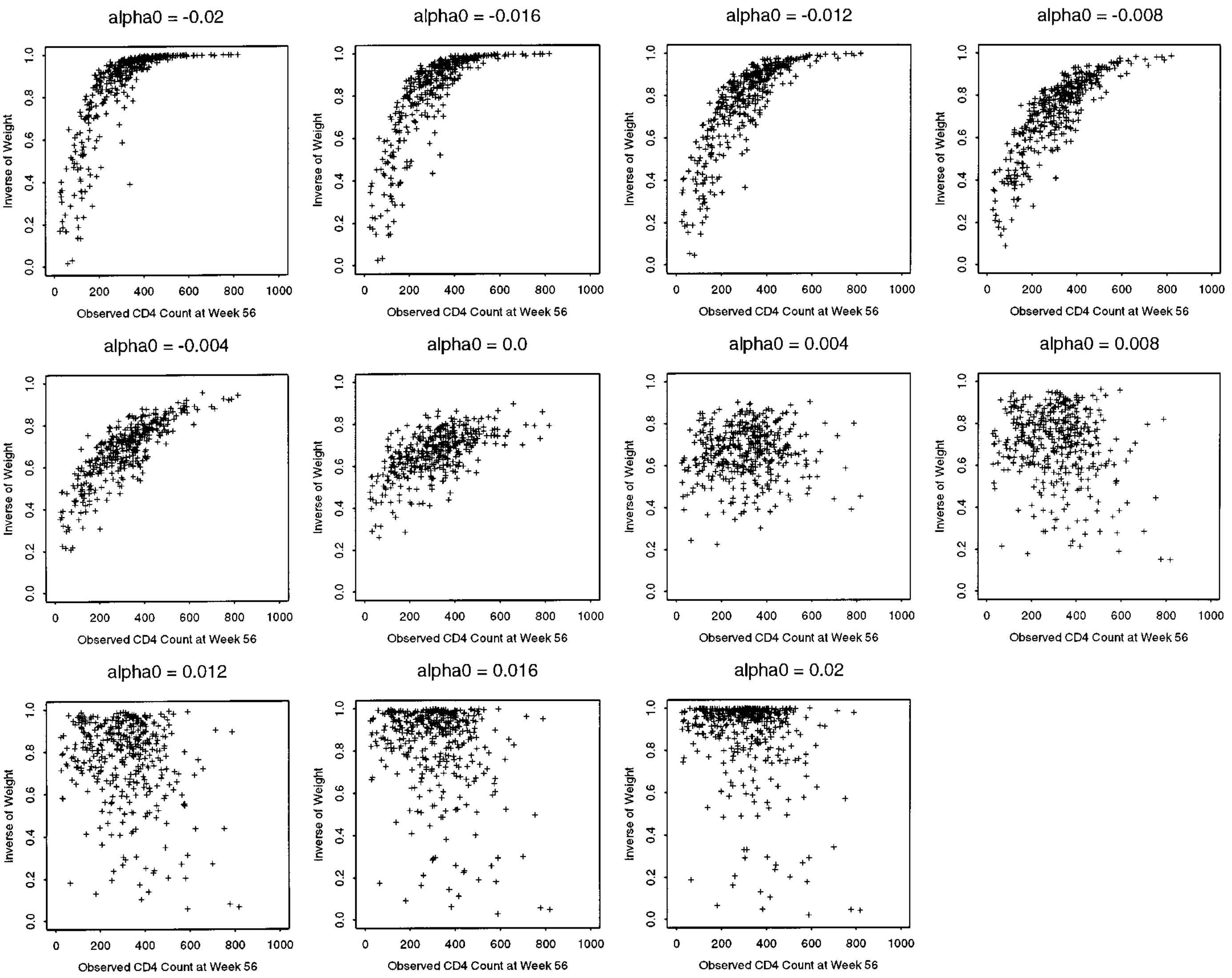

Figure 6. Observed CD4 at Week 56 Versus the Inverse of the Associated Estimated Weights in Model $B(\alpha_0)$.

α_0 is very positive, there is little correlation between the observed Y's and the estimated weights.

What are the lessons of this subsection? First, it is important to remember that the meaning of the selection bias parameter α_0 depends on the covariates $\bar{\mathbf{V}}(t)$ that are conditioned on in (1). Second, if we were previously so confused about how to logically map a range for α_{02} to one for α_{01} (or vice-versa), then how can we expect statistically naive subject matter experts to succeed at doing so? Obviously we cannot without providing guidance. Third, suggestion (d) of the preceding subsection may help avoid the breakdown of AIPCW estimators in model $A_1(\alpha_{01})$ in the sense that the value of α_{02} at which breakdown will occur will exceed, in absolute value, that of α_{01}. However, the foregoing discussion implies that this help would best be understood as possibly alerting our expert to the fact that her plausible range for α_{01} should not have reached the breakdown point after all.

7.2.4 Multiplicative Versus Additive Hazard Models. We argued earlier that the length of any plausible range for α_{02} should logically exceed that for α_{01} when the covariates $\bar{\mathbf{V}}(t)$ in model A_2 are highly associated with the outcome Y among the completers ($\Delta = 1$). Thus if one wishes to assess the relationship between plausible ranges for α_{01} and α_{02} before seeing the data and learning about F_O, then it is important to understand when $\bar{\mathbf{V}}(t)$ and Y will be highly associated among the completers. Subject matter experts will generally have stronger prior opinions about qualitative aspects of the marginal distribution of Y and $\bar{\mathbf{V}}(t)$ than about their conditional distribution given $\Delta = 1$. Thus it is important to recognize that when $\alpha_{02} \neq 0$ and $\bar{\mathbf{V}}(t)$ is a strong predictor of drop-out, Y and $\bar{\mathbf{V}}(t)$ can, in principle, be highly correlated given $\Delta = 1$ under the multiplicative hazard model (1), even if they are marginally independent. In contrast, if we had replaced the multiplicative hazard model A of (1) with the additive hazard model A_{add} that specifies

$$\lambda_Q(t|\bar{\mathbf{V}}(T), Y) = \lambda_0(t|\bar{\mathbf{V}}(t)) + \exp(r(t, \alpha_0; Y))$$

with $r(t, \alpha_0; Y)$ a known function, then marginal independence of $\bar{\mathbf{V}}(t)$ and Y imply conditional independence given $\Delta = 1$. Estimators of the mean μ_0 under an additive hazard model can be straightforwardly obtained by merging the methods described by Lin and Ying (1994) with the inverse probability of censoring weighting method described in Sections 3 and 4.

To clarify the potential importance of the foregoing observations, consider the following hypothetical scenario, where, to simplify matters, we now let models A_1 and $A_{\text{add},1}$ represent models in which no covariate data are available. Models A_2 and $A_{\text{add},2}$ denote models in which there is a single continuous covariate $\mathbf{V}$. Suppose that a subject matter expert believes that there is selection bias and specifies a plausible range of (.003, .010) for α_{01}. Data now become available on the covariate $\mathbf{V}$ that is known to be highly correlated with time to drop-out Q, and the data are reanalyzed using model $A_2(\alpha_{02})$. The expert is thus asked to provide a plausible range for α_{02}. Suppose that the expert is quite uncertain as to the marginal association of $\mathbf{V}$ and Y. At one extreme, he or she believes that it is possible that the assumed correlation between Q and Y (encoded in α_{01}) is completely explained by the covariate $\mathbf{V}$, and thus $\alpha_{02} = 0$. At the other extreme, he or she believes it is possible that $\mathbf{V}$ and Y are marginally independent. He or she thus provides a plausible interval (0, .010) for α_{02}, naively assuming that if $\mathbf{V}$ is independent of Y, then the magnitude of the conditional selection bias (encoded in α_{02}) will equal the magnitude of the marginal selection bias encoded in α_{01}. But the expert can be quite wrong. To see why, we consider a result of Hougaard (1986), which shows that if $\lambda_0(t|\mathbf{V}) = \lambda\mathbf{V}$ and $\mathbf{V}$ has an α-stable distribution with $\alpha < 1$, then $\alpha_{02} = \alpha_{01}/\alpha$. For instance, if it was known that α was .5, then logically the expert should have provided an interval of (0, .020) for α_{02} to be consistent with her interval for α_{01}. This reflects the fact that under these conditions, $\mathbf{V}$ and Y will be dependent given $\Delta = 1$. Thus even conditioning on a covariate $\mathbf{V}$ independent of Y can greatly increase the plausible range for the selection bias parameter. In contrast, under the additive hazard model, the expert would have been correct in his or her intuition that if $\mathbf{V}$ is independent of Y, then the marginal and conditional selection bias parameters α_{01} and α_{02} would be equal.

Should the foregoing counterintuitive result for the multiplicative hazard model $A(\alpha_0)$ suggest that we use the additive hazard model $A_{\text{add}}(\alpha_0)$ in conducting sensitivity analyses? We think not, for several reasons. First, epidemiologists are more familiar with modelling the shape and magnitude of a rate ratio function than a rate difference function. Second, marginal independence of Y and $\mathbf{V}$ does not imply conditional independence given $\Delta = 1$ even under an additive hazard model, when there is an interaction between Y and $\mathbf{V}$ on an additive hazard scale; that is, when we replace $r(t, \alpha_0; Y)$ by $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ in model A_{add}. Third, α-stable distributions with $\alpha < 1$ are quite pathological (e.g., they have no moments), and it may be rare to find Y and $\mathbf{V}$ strongly conditionally correlated when they are marginally independent.

7.3 Alternative Nonparametric Identified Models

Our NPI selection model $A(\alpha_0)$ is but one of many alternative NPI models that we could have used to conduct a sensitivity analysis. The choice among NPI models should ultimately depend on the ease with which subject matter experts can provide meaningful opinions about the sign and magnitude of the nonidentified selection bias parameters, such as the parameter α_0 in model $A(\alpha_0)$. In the following subsections we describe several alternative models and compare the strengths and weaknesses of the NPI model $A(\alpha_0)$ with those of the alternative models.

7.3.1 A Nonparametric Identified Selection Model for the Effect of Y Only on Selection. The most general form of our model $A(\alpha_0)$ is

$$\lambda_Q(t|\bar{\mathbf{V}}(T), Y) = \lambda_0(t|\bar{\mathbf{V}}(t)) \exp(r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)) \qquad (13)$$

with $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ known to the data analyst. Our model (1) specified that

$$r(t, \alpha_0; \bar{\mathbf{V}}(T), Y) = \alpha_0 Y. \qquad (14)$$

In Appendix A we prove that there can never be any data evidence contradicting (14), as $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ is not identified. Nonetheless, a subject matter expert would most likely be unwilling to believe (14), because there is no good substantive reason why drop-out at time t should depend on the unobserved future $(Y, \underline{\mathbf{V}}(t) \equiv \{\mathbf{V}(u); t \leq u \leq T\})$ only through the outcome of interest Y. Yet it seems an impossible burden for a subject matter expert to specify plausible functional forms for $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)$ if $\bar{\mathbf{V}}(T)$ is a high-dimensional process. Thus model $A(\alpha_0)$ may not be useful for a sensitivity analysis.

The way out of this apparent dilemma is to ask the subject matter expert to provide plausible functional forms only for the effect of the outcome of interest Y on drop-out at t, ignoring the future covariate process $\underline{\mathbf{V}}(t)$. Formally, this means that we consider a model $A^*(\alpha_0)$,

$$\lambda_Q(t|\bar{\mathbf{V}}(t), Y) = \lambda_0(t|\bar{\mathbf{V}}(t)) \exp(r^*(t, \alpha_0; \bar{\mathbf{V}}(t), Y)), \qquad (15)$$

for the hazard of drop-out at t conditional on and only on past covariate history $\bar{\mathbf{V}}(t)$ and future Y with $r^*(t, \alpha_0; \bar{\mathbf{V}}(t), Y)$ a known function satisfying $r^*(t, 0; \bar{\mathbf{V}}(t), Y) = 0, \lambda_0(t|\bar{\mathbf{V}}(t))$ an unrestricted positive function, and the parameter α_0 assumed known. Robins and Rotnitzky (1992) studied model $A^*(\alpha_0)$ in the special case in which $\alpha_0 = 0$. In Lemma A.1 of Appendix A, we show that model $A^*(\alpha_0)$ is a nonparametric model for the distribution F_O of the observed data. Furthermore, we show that the distribution of Y is identified under model $A^*(\alpha_0)$ (although the distribution of $\bar{\mathbf{V}}(T)$ is not identified). Under model $A^*(\alpha_0)$, the functional form $\alpha_0 Y$ for $r^*(t, \alpha_0; \bar{\mathbf{V}}(t), Y)$ might well be viewed as substantively plausible by a subject matter expert, because it only says that the effect of Y on the hazard of drop-out at time t has an exponential dependence on Y with no interaction on a hazard ratio scale with the past covariates $\bar{\mathbf{V}}(t)$. Model $B^*(\alpha_0)$, which imposes restriction (2) in addition to (15), may be used in lieu of $A^*(\alpha_0)$ when the covariate process $\bar{\mathbf{V}}(t)$ is high dimensional.

If we agree that when our goal is to make inferences concerning a functional of the marginal distribution of Y such as the mean μ_0, it is more natural to consider models $A^*(\alpha_0)$

and $B^*(\alpha_0)$ than models $A(\alpha_0)$ and $B(\alpha_0)$, then we are left with the question of how to estimate μ_0 under these new models. Fortunately, we can borrow, without modification, the estimation methods used for model $A(\alpha_0)$ and $B(\alpha_0)$ discussed in Sections 3 and 4. Specifically, suppose that we replace (1) by (15) with $r^*(t,\alpha_0;\bar{\mathbf{V}}(t),Y)=\alpha_0 Y$. Then the estimators $\hat{\mu}(b)$ and $\hat{\boldsymbol{\psi}}(\mathbf{b})$ of Sections 3 and 4 remain RAL estimators of μ_0 and $\boldsymbol{\psi}_0$. It may help to restate this result in a slightly different manner. Suppose that model $A(\alpha_0)$ given by (1) was misspecified, because $\lambda_Q(t|\bar{\mathbf{V}}(T),Y)$ actually depended on future covariate history $\underline{\mathbf{V}}(t)$. Suppose, however, that (15) was true with $r^*(t,\alpha_0;\bar{\mathbf{V}}(t),Y)=\alpha_0 Y$. Then the estimators of μ_0 in Sections 3 and 4 remain RAL estimators.

The foregoing results follow from the fact that, as shown in Lemma A.1, μ_0 is the same functional of F_O in both model $A(\alpha_0)$ and model $A^*(\alpha_0)$ when the function $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)$ specified in model $A(\alpha_0)$ does not depend on $\underline{\mathbf{V}}(t)$ and is equal to the function $r^*(t,\alpha_0;\bar{\mathbf{V}}(t),Y)$ specified in model $A^*(\alpha_0)$.

7.3.2 Nonparametric Identified Mean Models. Consider the "mean" model $A_{\text{mean}}(\alpha_0)$, which specifies that

$$\begin{aligned}&\Phi(E[Y|Q=t,\bar{\mathbf{V}}(t)])\\&=\Phi(E[Y|Q>t,\bar{\mathbf{V}}(t)])+r(t,\alpha_0;\bar{\mathbf{V}}(t)),\quad t\in(0,T),\end{aligned}\qquad(16)$$

where $r(t,\alpha_0;\bar{\mathbf{V}}(t))$ is a known selection bias function satisfying $r(t,0;\bar{\mathbf{v}}(t))=0;\alpha_0$ is a parameter assumed known to the data analyst, $\Phi(x)$ is a known monotone increasing function, and $E[Y|Q=t,\bar{\mathbf{V}}(t)]$ is assumed smooth in $(t,\bar{\mathbf{V}}(t))$. The function $r(t,\alpha_0;\bar{\mathbf{V}}(t))$ contrasts the mean of Y among subjects who drop out at time t with the mean among subjects continuing on study at t. Note that $\alpha_0=0$ implies the absence of selection bias on unobservables.

Suppose that Y is a dichotomous (0,1) variable and we choose $\Phi(x)=\ln(x/(1-x))$. Then it follows by an application of Bayes' theorem that model $A_{\text{mean}}(\alpha_0)$ with selection bias function $r(t,\alpha_0;\bar{\mathbf{v}}(t))$ is equivalent to model $A^*(\alpha_0)$ with selection bias function $r^*(t,\alpha_0;\bar{\mathbf{v}}(t),y)$ given by $r^*(t,\alpha_0;\bar{\mathbf{v}}(t),y)=yr(t,\alpha_0;\bar{\mathbf{v}}(t))$. In a modification of the usual nomenclature, in this case we refer to model $A_{\text{mean}}(\alpha_0)$ as a "continuous-time sequential-pattern-mixture" representation of the selection model $A^*(\alpha_0)$. Conversely, model $A^*(\alpha_0)$ with selection bias function $r^*(t,\alpha_0;\bar{\mathbf{v}}(t),y)$ is equivalent to model $A_{\text{mean}}(\alpha_0)$ with $r(t,\alpha_0;\bar{\mathbf{v}}(t))=r^*(t,\alpha_0;\bar{\mathbf{v}}(t),1)-r^*(t,\alpha_0;\bar{\mathbf{v}}(t),0)$. Having two representations may help subject matter experts in specifying functional forms for r and r^* and plausible values for the parameter α_0.

Next, suppose that Y is a random variable whose distribution may be arbitrary. Then Robins et al. (1999) showed that model $A_{\text{mean}}(\alpha_0)$ with $\Phi(x)=x$ is a nonparametric model for the distribution of F_O of the observed data. Furthermore, the marginal mean μ_0 of Y is identified via

$$\mu_0=E\left[\frac{\Delta\{Y+\int_0^T r(t,\alpha_0;\bar{\mathbf{V}}(t))\lambda_Q(t|\bar{\mathbf{V}}(t))\,dt\}}{S(T|\bar{\mathbf{V}}(T))}\right],\qquad(17)$$

where now $S(t|\bar{\mathbf{V}}(t))=\exp(-\int_0^t\lambda_Q(u|\bar{\mathbf{V}}(u))\,du)$ and $\lambda_Q(t|\bar{\mathbf{V}}(t))$ is the hazard of drop-out at t given $\bar{\mathbf{V}}(t)$. Robins et al. (1995) showed that when $\alpha_0=0$, the mean μ_0 of Y is the same functional of the law of F_O under the mean model $A_{\text{mean}}(\alpha_0)$ as under model $A^*(\alpha_0)$ and model $A(\alpha_0)$. When $\alpha_0\neq 0$, this equivalence no longer holds. Indeed, when $\alpha_0\neq 0$, in contrast to model $A^*(\alpha_0),\mu_0$ remains identified under model $A_{\text{mean}}(\alpha_0)$ even when the support of Y for the drop-outs $(\Delta=0)$ is not contained within the support of Y for the completers $(\Delta=1)$.

When $\bar{\mathbf{V}}(t)$ is high dimensional, we consider a model $B_{\text{mean}}(\alpha_0)$ that imposes, in addition to (16), the time-dependent Cox model $\lambda_Q(t|\bar{\mathbf{V}}(t))=\lambda_0(t)\exp(\theta_0'\mathbf{W}^*(t))$, where $\lambda_0(t)$ is an unspecified positive function, θ_0 is an unknown parameter to be estimated, and $\mathbf{W}^*(t)$ is a known vector-valued function of $\bar{\mathbf{V}}(t)$. We then estimate μ_0 by replacing the expectation in (17) by a sample average and estimating $\lambda_Q(t|\bar{\mathbf{V}}(t))$ and $S(t|\bar{\mathbf{V}}(t))$ based on the fit of the Cox model.

Finally, consider the case where Y is a positive random variable whose distribution is otherwise unrestricted. Then Robins et al. (1999) showed that model $A_{\text{mean}}(\alpha_0)$ with $\Phi(x)=\ln(x)$ is a nonparametric model for F_O and that the marginal mean μ_0 of Y is identified. These authors also provided an identifying formula for μ_0 and proposed a method of estimation.

7.3.3 Comparison of Model $\boldsymbol{A}^*(\boldsymbol{\alpha}_0)$ With $\boldsymbol{A}_{\text{mean}}(\boldsymbol{\alpha}_0)$. We spoke with a number of epidemiologists about our sensitivity analysis of ACTG 175. They were split as to the ease with which they could provide meaningful opinions about the selection bias functions and parameters in model $A^*(\alpha_0)$ versus model $A_{\text{mean}}(\alpha_0)$ with $\Phi(x)$ either x or $\ln(x)$. Those who preferred model $A_{\text{mean}}(\alpha_0)$ made two points. First, the mean model, in contrast to the selection model, asks one to form opinions about unknowns that are directly related to the final estimated of interest, the mean μ_0 of Y. (However, an advantage of the selection model $A^*(\alpha_0)$ over the mean model is that one can treat the multiple functionals of the law of Y that may be of interest in a unified fashion.) Second, because the outcome Y is observed only (if ever) at week 56, it is more natural to think about the mean of Y given the past as in model $A_{\text{mean}}(\alpha_0)$ than to think about the effect of the yet (and possibly never) to be observed Y on drop out at earlier times as in model $A^*(\alpha_0)$. This criticism of model $A^*(\alpha_0)$ loses some of its sting if we allow our experts to reassess their plausible range for α_0 after the data have been analyzed and they are provided with both the difference between the mean of Y in the completers and the estimated (i.e., imputed) mean in the drop-outs and a plot of the estimated weights as a function of Y. At first, this might seem totally unacceptable. However, as pointed out by I. J. Good (Good 1983), if the expert were to carry out this reassessment for multiple simulated datasets before seeing the actual data, then this would be a perfectly valid method for eliciting the expert's actual prior beliefs about α_0. This method is Good's "device of imaginary results." If one's prior uncertainty concerning the distribution F_O of the observed data is much less than that concerning the nonidentified selection bias parameter α_0, as should often be the case, then most of the "imaginary" (i.e., simulated) datasets would result in a similar "reassessed" plausible range for α_0. In this case one perhaps might dispense with the simulations altogether and reassess only the actual data. The problem of course is that we may, as humans, react quite differently to the same data, depending on whether we know it to be real versus "imaginary."

A major difference between models $A^*(\alpha_0)$ and $A_{\text{mean}}(\alpha_0)$ is that the latter model can lead to out-of-sample extrapolation. Specifically, as discussed in Section 7.2, the AIPCW estimators $\hat{\mu}(b)$ of μ_0 in model $A^*(\alpha_0)$ are guaranteed to lie within the range of the observed Y's. In contrast, for large values of α_0, estimates of μ_0 under model $A^*_{\text{mean}}(\alpha_0)$ can lie outside that range. In general, a method that can extrapolate far outside the range of the data will be extremely sensitive to the choice of selection bias function and should be used with caution. However, whenever it is considered plausible that selection bias is so extreme that the support of Y in the dropouts and the completers differ, it is essential to use methods that extrapolate.

8. SUMMARY

In this article we have shown how to estimate the mean μ_0 of an outcome of interest Y measured at a fixed time T when (a) some subjects drop out of the study, (b) drop out is nonignorable, (c) the probability of nonresponse is a function of Y and additional time-independent and time-dependent covariates $\bar{\mathbf{V}}(t)$ and follows a semiparametric model, and (d) no restrictions are placed on the joint distribution of $\bar{\mathbf{V}}(T)$ and Y. From a practical and philosophical perspective, we argued that it was more natural to fix the parameter associated with Y in the nonresponse model and perform a sensitivity analysis to see how inference about μ_0 changes as we vary this parameter over a plausible range of values. We illustrated our technique with data from an AIDS clinical trial, ACTG 175. We described settings in which our method breaks down and offered alternative methods appropriate for these settings. We discussed, but left unresolved, the question of whether subject matter experts will be able to provide a plausible range of values for the selection bias parameter. If not, then these methods may ultimately be of limited scientific value. On a similar note, it is unknown whether our formal methods will prove of more use to practicing scientists than the informal sensitivity analyses they already conduct, based on "back of the envelope" calculations. Our guess is that many scientists underestimate uncertainty, and our formal methods combined with informative graphical displays can usefully serve as a brusque reminder of just how much is uncertain.

In Appendix B we present a general theory for constructing estimators of a parameter of interest when drop-out is nonignorable and both the full-data and nonresponse mechanism follow semiparametric models. In later reports we plan to use this general theory to extend our results to failure time outcomes, repeated-measures outcomes, and regression and counterfactual causal models for the effect of baseline and time-dependent covariates on the outcome. As we have worked only with the proportional hazards nonresponse models, we also plan to develop methods to handle other nonresponse models, such as the accelerated failure time and additive hazard models.

APPENDIX A: PROOF OF THEOREM 1

We actually prove a generalization of Theorem 1 in which we replace (1) by the more general expression (13), where $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)$ is any known function satisfying condition 3 given here. Throughout this proof, we assume that the observed data law F_O is absolutely continuous with respect to a dominating measure ν and we denote expectations, densities, and probabilities under F_O with the superscript *. We assume the following conditions:

1. Given $\Delta = 1$, the process $\bar{\mathbf{V}}(T)$ has CADLAG sample paths with all discontinuities occurring at a finite number of fixed nonrandom times $0 \equiv t_0 < t_1 < \cdots < t_M$ with $t_M < T$.
2. For $t \in [0,T), \lambda^*_Q(t|\bar{\mathbf{V}}(t))$ is bounded by a constant c with probability 1 and has a bounded derivative except at $t_k, k \in \{0,\ldots,M\}$.
3. For $t \in [0,T)$ and $\bar{\mathbf{v}}(t)$ in the support of $\bar{\mathbf{V}}(t)$ on $\Delta = 1, |r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)| < k(\alpha_0)$ with probability 1 for some constant $k(\alpha_0)$ under $F^*_{\bar{\mathbf{V}}(T),Y|\Delta=1,\bar{\mathbf{V}}(t)=\bar{\mathbf{v}}(t)}$, and $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)$ has a bounded derivative with respect to t except at $t_k, k \in \{0,\ldots,M\}$.
4. $f^*(Y|\Delta = 1,\bar{\mathbf{V}}(t_M))$ and $f^*(\mathbf{V}(t_k)|\bar{\mathbf{V}}(t_{k-1}), Q > t_k)$ are bounded with probability 1 for $k \in \{0,\ldots,M\}$.

We first prove part (b) of the theorem. We establish that the map from (F_O,α_0) to $F_{\bar{\mathbf{V}}(T),Y}$ and $\lambda_0(t|\bar{\mathbf{V}}(t))$ is given by

$$f(\bar{\mathbf{v}}(T),y) = g(0,y,\bar{\mathbf{v}}(T))f^*(\bar{\mathbf{v}}(0)) \tag{A.1a}$$

and

$$\lambda_0(t|\bar{\mathbf{v}}(t)) = \lambda^*_Q(t|\bar{\mathbf{v}}(t))/p_g(t,\alpha_0,\bar{\mathbf{v}}(t)), \tag{A.1b}$$

where

$$p_g(t,\alpha_0,\bar{\mathbf{v}}(t)) = \int\int \exp(r(t,\alpha_0;\bar{\mathbf{v}}(T),y)) \times g(t,y,\bar{\mathbf{v}}(T))\,d\nu(y) \prod_{k:t_k>t} d\nu(\mathbf{v}(t_k))$$

and $g(t,y,\bar{\mathbf{v}}(T))$ is the unique solution on $t \in [0,T]$ to the nonlinear Volterra integral equation

$$\begin{aligned} g(t,y,\bar{\mathbf{v}}(T)) &= f^*(y|\Delta = 1,\bar{\mathbf{v}}(t_M)) \prod_{k:t_k>t} f^*(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}), Q > t_k) \\ &\quad \times \exp\left(-\int_t^T \lambda^*_Q(x|\bar{\mathbf{v}}(x)) \left\{1 - \frac{\exp(r(x,\alpha_0;\bar{\mathbf{v}}(T),y))}{p_g(x,\alpha_0,\bar{\mathbf{v}}(x))}\right\} dx\right). \end{aligned} \tag{A.1c}$$

We now prove (A.1a)–(A.1c). First, it can be shown that, under Assumption 2, the conditional densities on the right side of (A.1c) have well-defined versions that are continuous (with respect to the weak topology) in all of their arguments, by arguing as in Gill and Robins (1999). Assumptions 3 and 4 guarantee that $\exp(r(x,\alpha_0;\bar{\mathbf{v}}(T),y))/p_g(x,\alpha_0,\bar{\mathbf{v}}(x))$ in (A.1c) is bounded. Arguing as in the work of Tricomi (1957, sec. 1.13), it follows that by the smoothness assumptions 2–4, (A.1c) has a solution, and it is unique. We next show that the right side of (A.1a) integrates to 1, so $f(\bar{\mathbf{v}}(T),y)$ is a density. It is sufficient to show that $Z(t) \equiv \int\int g(t,y,\bar{\mathbf{v}}(T))\,d\nu(y)\prod_{k:t_k>t} d\nu(\mathbf{v}(t_k)) = 1$ for all t. Under our smoothness assumptions, $Z(t)$ is continuous on $[0,T]$ and differentiable except at the t_k's. Now it follows from (A.1c) and the definition of p_g that $\dot{Z}(t)=\lambda_Q^*(t|\bar{\mathbf{v}}(t))(Z(t)-1)$ for $t\neq t_k$, where for any function $h(t,\cdot),\dot{h}(t,\cdot)=\partial h(t,\cdot)/\partial t$. Because $Z(t)$ is equal to 1 at $t=T$, it follows by the uniqueness of solutions to differential equations that $Z(t)=1$ for all t.

We now prove that any candidate law $F_{\bar{\mathbf{V}}(T),Y,Q}$ marginalizing to F_O satisfying (13) will have density $f(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)$ that satisfies (A.1c) and $\lambda_0(t|\bar{\mathbf{v}}(t))$ given by (A.1b). If $F_O\equiv F^*$ is the marginal of our candidate law, then

$$\begin{aligned} f(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)\Pr[\Delta=1|\bar{\mathbf{v}}(T),y,Q>t]& \\ = f^*(\Delta=1,\bar{\mathbf{v}}(T),y|\bar{\mathbf{v}}(t),Q>t).& \qquad \text{(A.2)}\end{aligned}$$

Dividing both sides of (A.2) by $\Pr[\Delta=1|\bar{\mathbf{v}}(T),y,Q>t]=\exp(-\int_t^T\lambda_Q(x|\bar{\mathbf{v}}(T),y)\,dx)$ and noting that

$$\begin{aligned} &f^*(\Delta=1,\bar{\mathbf{v}}(T),y|\bar{\mathbf{v}}(t),Q>t) \\ &\quad = f^*(y|\Delta=1,\bar{\mathbf{v}}(t_M))\prod_{k:t_k>t}f^*(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}),Q>t_k) \\ &\qquad \times\exp\left(-\int_t^T\lambda_Q^*(x|\bar{\mathbf{v}}(x))\right), \qquad \text{(A.3)}\end{aligned}$$

and, by (13),

$$\frac{\lambda_Q(t|\bar{\mathbf{v}}(T),y)}{\lambda_Q^*(t|\bar{\mathbf{v}}(t))}=\frac{\exp(r(t,\alpha_0;\bar{\mathbf{v}}(T),y))}{p_f(t,\alpha_0,\bar{\mathbf{v}}(t))}, \qquad \text{(A.4)}$$

where f in p_f is equal to $f(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)$, we obtain that $f(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)$ must solve (A.1c). It thus follows that (A.1a) is the unique candidate density $f(y,\bar{\mathbf{v}}(T))$ for the marginal law of $(Y,\bar{\mathbf{V}}(T))$. Further, (A.4) and (13) imply the unique candidate (A.1b) for $\lambda_0(t|\bar{\mathbf{v}}(t))$. Thus it only remains to show that our unique candidate law $F_{\text{cand}}\equiv F_{\bar{\mathbf{V}}(T),Y,Q}$ determined by (A.1a)–(A.1c) has marginal law F_{marg} for O equal to the given F^*.

Now it is easy to see by inspection that the map from F^* to the solution g of (A.1c) is 1 to 1. Hence we can conclude that $F_{\text{marg}}=F^*$ if we can show that the density $f_{\text{cand}}(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t,)$ derived from F_{cand} solves (A.1c), because, by the arguments in (A.2)–(A.4), this density solves (A.1c) when F^* is replaced by F_{marg} in (A.1c). Now, taking derivatives with respect to t for $t\neq t_k$, by (A.1c), $\dot{g}(t,y,\bar{\mathbf{v}}(T))=g(t,y,\bar{\mathbf{v}}(T))\lambda_0(t|\bar{\mathbf{v}}(t))\{p_g(t,\alpha_0,\bar{\mathbf{v}}(t))-\exp(r(t,\alpha_0;\bar{\mathbf{v}}(T),y))\}$. Also,

$$\begin{aligned} &\dot{f}_{\text{cand}}(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)=f_{\text{cand}}(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t) \\ &\qquad \times\lambda_0(t|\bar{\mathbf{v}}(t))\{p_{f_{\text{cand}}}(t,\alpha_0,\bar{\mathbf{v}}(t))-\exp(r(t,\alpha_0;\bar{\mathbf{v}}(T),y))\},\end{aligned}$$

because

$$\begin{aligned} &f_{\text{cand}}(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t) \\ &\quad = g(0,y,\bar{\mathbf{v}}(T))\exp\left\{-\int_0^t\lambda_0(x|\bar{\mathbf{v}}(x))\exp(r(x,\alpha_0;\bar{\mathbf{v}}(T),y))\right\} \\ &\quad \div\left\{\int\!\!\int\!\!\int g(0,y,\bar{\mathbf{v}}(T))\right. \\ &\qquad \times\exp\left\{-\int_0^t\lambda_0(x|\bar{\mathbf{v}}(x))\exp(r(x,\alpha_0;\bar{\mathbf{v}}(T),y))\right\} \\ &\qquad \left.\times\, d\nu(y)\prod_{k:t_k>t}d\nu(\mathbf{v}(t_k))\right\}.\end{aligned}$$

Further, from its definition, $\dot{f}_{\text{cand}}(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(0),Q>0)=g(0,y,\bar{\mathbf{v}}(T))$. Hence, by the uniqueness of solutions to differential equations, $\dot{f}_{\text{cand}}(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)=g(t,y,\bar{\mathbf{v}}(T))$ for $t\in[0,t_1)$. But then they are equal at t_1 and thus on $[t_1,t_2)$. Continuing, we conclude equality for all t, proving part (b) of the theorem.

The proof of part (a) follows from the fact that the foregoing construction depends on the values of α_0. As α_0 varies, the mapping from the law of the observed data to the law of the full data changes. This shows the lack of identifiability.

Corollary A.1. In model $A(\alpha_0)$ defined by (13), if $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)=r^*(t,\alpha_0;\bar{\mathbf{V}}(t),Y)$, then the density $f(y|\bar{\mathbf{v}}(t),Q>t)$ is the unique solution $g(t,y,\bar{\mathbf{v}}(t))$ to the nonlinear Volterra integral equation

$$\begin{aligned} &g(t,y,\bar{\mathbf{v}}(t)) \\ &\quad = \int\cdots\int\int f^*(y|\Delta=1,\bar{\mathbf{v}}(t_M)) \\ &\qquad \times\prod_{k:t_k>t}dF^*(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}),Q>t_k) \\ &\qquad \times\exp\left(-\int_t^T\lambda_Q^*(x|\bar{\mathbf{v}}(x))\right. \\ &\qquad \left.\times\left\{1-\frac{\exp(r^*(x,\alpha_0;\bar{\mathbf{v}}(x),y))}{\int\exp(r^*(x,\alpha_0;\bar{\mathbf{v}}(x),y))g(x,y,\bar{\mathbf{v}}(x))\,d\nu(y)}\right\}dx\right), \\ &\qquad\qquad \text{(A.5a)}\end{aligned}$$

and $\lambda_0(t|\bar{\mathbf{v}}(t))$ is given by

$$\begin{aligned} &\lambda_0(t|\bar{\mathbf{v}}(t))=\lambda_Q^*(t|\bar{\mathbf{v}}(t)) \\ &\qquad \div\int\exp(r^*(t,\alpha_0;\bar{\mathbf{v}}(t),y))f(y|\bar{\mathbf{v}}(t),Q>t)\,d\nu(y). \qquad \text{(A.5b)}\end{aligned}$$

Proof. Under our regularity conditions, (A.5a) has a unique solution. In the proof of Theorem 1 we showed that $f(y,\bar{\mathbf{v}}(T)|\bar{\mathbf{v}}(t),Q>t)$ solves (A.1c). The corollary now follows from (A.1b) and (A.1c) on integrating out the necessary $\mathbf{v}(t_k)$.

When $\alpha_0=0$, so $r^*(x,\alpha_0;\bar{\mathbf{v}}(x),y)=0$, the solution to (A.5a) is given by the g-computation algorithm formula of Robins (1986, 1987),

$$\begin{aligned} f(y|\bar{\mathbf{v}}(t),Q>t)&=\int\cdots\int f^*(y|\Delta=1,\bar{\mathbf{v}}(t_M)) \\ &\qquad \times\prod_{k:t_k>t}dF^*(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}),Q>t_k).\end{aligned}$$

Lemma A.1. Suppose that assumptions 1–4 of Theorem 1 hold, except that in 3 we replace r by r^* and T by t. Then model $A^*(\alpha_0)$ defined by (15) is a nonparametric model for F_O with $f(y|\bar{\mathbf{v}}(t),Q>t)$ identified and given by (A.5a) and $\lambda_0(t|\bar{\mathbf{v}}(t))$ given by (A.5b).

Proof. As in the proof of Theorem 1, let f^* denote densities under F_O. By the proof of Theorem 1, (A.6) has unique nonnegative solutions $g_1(t,y,\bar{\mathbf{v}}(t)),g_2(\bar{\mathbf{v}}(t_k),y),k\in\{k\colon t_k>t\}$, and $\rho(t,\bar{\mathbf{v}}(t))$ satisfying the unit integral restrictions $1=\int g_1(t,y,\bar{\mathbf{v}}(t))\,d\nu(y)=\int g_2(\bar{\mathbf{v}}(t_k),y)\,d\nu(\mathbf{v}(t_k))$:

$$\begin{aligned} &\{f^*(q,\bar{\mathbf{v}}(q),y|\bar{\mathbf{v}}(t),Q>t)\}^{\delta}\{f^*(q,\bar{\mathbf{v}}(q)|\bar{\mathbf{v}}(t),Q>t)\}^{1-\delta} \\ &\quad = \{\rho(q,\bar{\mathbf{v}}(q))\exp(r^*(q,\alpha_0;\bar{\mathbf{v}}(q),y))\}^{\delta} \\ &\quad \times\int(d\nu(y))^{1-\delta}g_1(t,y,\bar{\mathbf{v}}(t))\prod_{k:t<t_k<q}g_2(\bar{\mathbf{v}}(t_k),y)\end{aligned}$$

$$\times \exp\left(-\int_t^q \rho(x,\bar{\mathbf{v}}(x))\exp(r^*(x,\alpha_0;\bar{\mathbf{v}}(x),y))\,dx\right), \quad \text{(A.6)}$$

where if $\delta = 1$, then the first integration is not performed. As model $A(\alpha_0)$ is nonparametric, we can regard (A.6) as the density of the observed data given the event $(\bar{\mathbf{V}}(t) = \bar{\mathbf{v}}(t), Q > t)$ under model $A(\alpha_0)$ with $r(t,\alpha_0;\bar{\mathbf{v}}(T),y) = r^*(t,\alpha_0;\bar{\mathbf{v}}(t),y)$. Hence we conclude that $g_1(t,y,\bar{\mathbf{v}}(t)) = f(y|\bar{\mathbf{v}}(t),Q > t)$ solves (A.5a), $\rho(t,\bar{\mathbf{v}}(t)) = \lambda_0(t|\bar{\mathbf{v}}(t))$ is given by (A.5b), and $g_2(\bar{\mathbf{v}}(t_k),y) = f(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}),y) = f(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}),y,Q > t_k)$ is determined by the law $f_{\bar{\mathbf{V}}(T),Y}$ defined by (A.1a)–(A.1c), as (A.6) uniquely determines the conditional density of $(\bar{\mathbf{V}}(T),Y,Q)$ through the foregoing formulas, according to our proof of Theorem 1 and Corollary A.1. However, (A.6) is also the density of the observed data under model $A^*(\alpha_0)$ [i.e., (15)], with $g_2(\bar{\mathbf{v}}(t_k),y)$ now only $f(\mathbf{v}(t_k)|\bar{\mathbf{v}}(t_{k-1}),y,Q > t_k)$. Because we have just shown that (A.6) has a unique solution satisfying the foregoing positivity and unit integral restriction, we conclude that there exists only one law $F_{Y,Q,\bar{\mathbf{V}}(Q)|\bar{\mathbf{V}}(t),Q>t}$ that satisfies (15) and marginalizes to $F^*_{Q,\bar{\mathbf{V}}(Q),\Delta Y|Q>t,\bar{\mathbf{V}}(t)}$, and that this law has $\lambda_0(t|\bar{\mathbf{v}}(t))$ satisfying (A.5b) and $f(y|\bar{\mathbf{v}}(t),Q > t)$ solving (A.5a). This concludes the proof of the lemma.

APPENDIX B: GENERAL THEORY

Let L denote the complete (full) data. Suppose that we observe only $(R, L_{(R)})$, where $L_{(R)} = \varphi_R(L)$ and $\varphi_r(L)$ is a known function of L that depends on r. Specifically, R indexes the part of L that is actually observed. We assume that there exists a unique value of R, r^*, such that $\varphi_{r^*}(L) = L$. Let $\Delta = I(R = r^*)$. Furthermore, we assume that (a) L follows an arbitrary semiparametric model, F_L, indexed by a $p \times 1$ parameter $\boldsymbol{\mu}$ and an infinite-dimensional parameter $\boldsymbol{\theta}$; (b) R given L follows an arbitrary semiparametric model, $F_{R|L}$, indexed by a $q \times 1$ parameter $\boldsymbol{\gamma}$ and an infinite-dimensional parameter $\boldsymbol{\eta}$; and (c) $\Pr[\Delta = 1|L] > \sigma > 0$. We assume that the parameters in model F_L are variation independent of those in the model $F_{R|L}$. We let $\boldsymbol{\mu}_0, \boldsymbol{\gamma}_0, \boldsymbol{\theta}_0$, and $\boldsymbol{\eta}_0$ denote the true values of $\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\theta}$, and $\boldsymbol{\eta}$. We are interested in estimating $\boldsymbol{\psi}_0 = (\boldsymbol{\mu}_0', \boldsymbol{\gamma}_0')'$. We observe n independent and identically distributed copies of $O = (R, L_{(R)})$.

Let $\Lambda_1 = \Lambda(F_L)$ and $\Lambda_2 = \Lambda(F_{R|L})$ denote the (nuisance) tangent spaces for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ had we observed (R, L). (For a definition of nuisance tangent space see, e.g., Newey 1990). Throughout, our spaces are subspaces of the Hilbert space of $q+p$-dimensional mean 0 random vectors with the covariance inner product, and $\Pi(\cdot|\cdot)$ denotes the projection operator. Note that $\Lambda(F_L)$ and $\Lambda(F_{R|L})$ are orthogonal. For the "observed data," there is an induced semiparametric model that we denote by O. In model O the observed data nuisance tangent space is $\Lambda^O = \overline{\Lambda_1^O + \Lambda_2^O}$, where Λ_1^O is the observed data nuisance tangent space for $\boldsymbol{\theta}$ and Λ_2^O is the observed data nuisance tangent space for $\boldsymbol{\eta}$. Specifically, $\Lambda_j^O = \overline{R(g \circ \Pi_j)}$, where $R(\cdot)$ is the range of an operator, g: $\Omega^{(R,L)} \to \Omega^{(R,L_{(R)})}, g(\cdot) = E[\cdot|R, L_{(R)}], \Omega^{(R,L)}$ and $\Omega^{(R,L_{(R)})}$ are spaces of $p+q$-dimensional mean 0 random functions of (R, L) and $(R, L_{(R)}), \Pi_j$ is the projection operator from $\Omega^{(R,L)}$ onto Λ_j, and $\bar{\mathcal{S}}$ denotes the closed linear span of the set $\mathcal{S}$ (Bickel et al. 1993). A space superscripted by $\perp$ denotes the orthogonal complement of that space. We are interested in finding $\Lambda^{O,\perp}$ because in sufficiently smooth models including all those studied in this article, the set of influence functions of all RAL estimators of $\boldsymbol{\psi}_0$ is the set $\{E[\mathcal{A}\mathbf{S}_\psi']^{-1}\mathcal{A}; \mathcal{A} \in \Lambda^{O,\perp}\}$, where $\mathbf{S}_\psi$ is the observed data score for $\boldsymbol{\psi}$ evaluated at the truth. Another motivation for our interest in this space is as follows. An element in the $\Lambda^{O,\perp}$ space is a $(p+q)$-dimensional function of the observed data for an individual and the true values of the parameters $\boldsymbol{\psi}_0, \boldsymbol{\theta}_0$, and $\boldsymbol{\eta}_0$. Denote this function by $\mathbf{H} \equiv \mathbf{h}(O;\boldsymbol{\psi}_0,\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$. Suppose that we estimate $\boldsymbol{\psi}_0$ by $\hat{\boldsymbol{\psi}}$ solving $\sum_{i=1}^n \mathbf{h}(O_i;\boldsymbol{\psi},\hat{\boldsymbol{\theta}}(\boldsymbol{\psi}),\hat{\boldsymbol{\eta}}(\boldsymbol{\psi})) = 0$, where $\hat{\boldsymbol{\theta}}(\boldsymbol{\psi}_0)$ and $\hat{\boldsymbol{\eta}}(\boldsymbol{\psi}_0)$ converge to $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}_0$. Then Bickel et al. (1993) and Newey (1990) showed that under suitable regularity conditions $\hat{\boldsymbol{\psi}}$ is a RAL estimator with influence function $\boldsymbol{\tau}^{-1}\mathbf{H}$, where $\boldsymbol{\tau} = E[\mathbf{H}\mathbf{S}_\psi'] = -\partial E[\mathbf{h}(O;\boldsymbol{\psi},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)]/\partial\boldsymbol{\psi}_{|\boldsymbol{\psi}=\boldsymbol{\psi}_0}$. But this is the same influence function as would have been obtained by solving the estimating equation $\sum_{i=1}^n \mathbf{h}(O_i;\boldsymbol{\psi},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) = 0$ in which the infinite-dimensional components $(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)$ are known rather than estimated. It is precisely the orthogonality of $\mathbf{H}$ to Λ^O that obviated the need to adjust the asymptotic variance for estimation of the nuisance parameters.

Taking orthogonal complements, $\Lambda^{O,\perp} = \Lambda_1^{O,\perp} \cap \Lambda_2^{O,\perp}$. Let $\mathbf{a}(L)$ and $\mathbf{b}(R, L_{(R)})$ be $p+q$-dimensional functions of L and $(R, L_{(R)})$. Rotnitzky and Robins (1997) showed how to compute $\Lambda_1^{O,\perp}$. Specifically,

$$\Lambda_1^{O,\perp} = \{\Delta \Pr[\Delta = 1|L]^{-1}\mathbf{a}(L) + \mathbf{b}(R, L_{(R)}): \\ \mathbf{a}(L) \in \Lambda(F_L)^\perp \text{ and } E[\mathbf{b}(R, L_{(R)})|L] = 0\}.$$

By the relationship between ranges and null spaces, we know that $\Lambda_2^{O,\perp} = N(\Pi_2^T \circ g^T)$, where $N(\cdot)$ is the null space of an operator and superscript T denotes the adjoint of an operator. As a projection operator, $\Pi_2^T = \Pi_2$ and g^T is the identity operator. So,

$$\begin{aligned}\Lambda_2^{O,\perp} &= \{\mathbf{b}(R, L_{(R)}): \Pi[\mathbf{b}(R, L_{(R)})|\Lambda(F_{R|L})] = 0\} \\ &= \{\mathbf{b}(R, L_{(R)}): \mathbf{b}(R, L_{(R)}) \in \Lambda(F_{R|L})^\perp\}.\end{aligned}$$

B.1. Application of the General Theory to Models $A(\alpha_0)$ and $B(\alpha_0)$

We now apply this general theory to obtain $\Lambda^{O,\perp}$ in models $A(\alpha_0)$ and $B(\alpha_0)$. In these models we have $L = (\bar{\mathbf{V}}(T), Y), R = Q, L_{(R)} = \varphi_Q(L) = (\bar{\mathbf{V}}(Q), I(Q = T)Y), r^* = T$, and $\Delta = I(R = T)$. We actually consider a generalization of model $A(\alpha_0)$ in which (1) is replaced with the more general expression (13), where $r(t,\alpha_0;\bar{\mathbf{V}}(T),Y)$ is any known function satisfying the condition 3 of Theorem 1 in Appendix A. Define $S(t) = \exp(-\int_0^t \lambda_0(u|\bar{\mathbf{V}}(u))\exp(r(u,\alpha_0;\bar{\mathbf{V}}(T),Y))\,du), S \equiv S(T)$, and $\varepsilon = Y - \mu_0$. Note that under (13), $S = P[\Delta = 1|\bar{\mathbf{V}}(T),Y]$. Note also that under model $B(\alpha_0), \lambda_0(t|\bar{\mathbf{V}}(t)) = \lambda_0(t)\exp\{\boldsymbol{\gamma}_0'\mathbf{W}(t)\}$, and hence $S(t)$ depends on the value of $\boldsymbol{\gamma}_0$. To emphasize this dependence, when developing the results for model $B(\alpha_0)$, we write $S(t;\boldsymbol{\gamma}_0)$ for $S(t)$. Throughout, we use $N(t)$ to denote the counting process for censoring $I(Q \le t, \Delta = 0)$ and $M(t) = N(t) - \int_0^t I(Q \ge u)\lambda_0(u|\bar{\mathbf{V}}(u))\exp\{r(u,\alpha_0;\bar{\mathbf{V}}(T),Y)\}\,du$ to denote its associated martingale.

In model $A(\alpha_0)$,

$$\begin{aligned}\Lambda(F_L) &= \{a(\varepsilon) + a(\bar{\mathbf{V}}(T),\varepsilon): a(\varepsilon) \text{ and } a(\bar{V}(T),\varepsilon) \\ &\qquad \text{are scalar functions with } E[a(\varepsilon)] = E[\varepsilon a(\varepsilon)] \\ &= E[a(\bar{V}(T),\varepsilon)|Y] = 0\}\end{aligned}$$

and

$$\Lambda(F_L)^\perp = \{k\varepsilon: k \in R^1\}$$

by theorem 8.3 of Robins et al. (1994). To compute $\Lambda_1^{O,\perp}$, note that any function $c(R, L_{(R)})$ admits the unique representation $\Delta a(\bar{\mathbf{V}}(T),Y) + (1-\Delta)b(\bar{\mathbf{V}}(Q),Q)$. Thus $E[c(R, L_{(R)})|L] = 0$ if and only if

$$c(R, L_{(R)}) = -\Delta \frac{E[(1-\Delta) b(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y]}{S} + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q).$$

Thus we have

$$\Lambda_1^{O,\perp} = \left\{ \Delta \left(\frac{k\varepsilon}{S} - \frac{E[(1-\Delta) b(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y]}{S} \right) + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q) \colon \ k \in R^1 \right\}.$$

In model $B(\alpha_0)$ the preceding representations of $\Lambda(F_L), \Lambda(F_L)^\perp$, and $\Lambda_1^{O,\perp}$ still hold, except that k is replaced by $\mathbf{k} \in R^{q+1}$, and the functions a and b are $(q+1)$ dimensional and denoted by $\mathbf{a}$ and $\mathbf{b}$.

B.1.1 Computation of $\Lambda^{O,\perp}$ in Model $A(\alpha_0)$. To compute $\Lambda^{O,\perp}$, we first compute $\Lambda_2^{O,\perp}$, then intersect it with $\Lambda_1^{O,\perp}$ given earlier. It follows from Ritov and Wellner (1988) that in model $A(\alpha_0)$,

$$\Lambda(F_{R|L}) = \left\{ \int_0^T g(t, \bar{\mathbf{V}}(t)) \, dM(t) \colon \ g(t, \bar{\mathbf{V}}(t)) \text{ is an arbitrary function of } t \text{ and } \bar{\mathbf{V}}(t) \right\}$$

and

$$\Lambda(F_{R|L})^\perp = \left\{ a(\bar{\mathbf{V}}(T), Y) + \int_0^T g(t, \bar{\mathbf{V}}(T), Y) \, dM(t) \colon \ E[g(t, \bar{\mathbf{V}}(T), Y) | Q = t, \bar{\mathbf{V}}(t)] = 0, \text{ and } E[a(\bar{\mathbf{V}}(T), Y)] = 0 \right\}.$$

To compute $\Lambda_2^{O,\perp}$, we write an arbitrary function $c(R, L_{(R)}) \equiv \Delta a(\bar{\mathbf{V}}(T), Y) + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q)$ as

$$\begin{aligned} c(R, L_{(R)}) &= \left\{ \Delta A + (1-\Delta) B + \frac{\Delta}{S} E[(1-\Delta) B | \bar{\mathbf{V}}(T), Y] \right\} \\ &\quad - \frac{\Delta}{S} E[(1-\Delta) B | \bar{\mathbf{V}}(T), Y] \\ &= \frac{\Delta}{S} m(\bar{\mathbf{V}}(T), Y) + (1-\Delta) B \\ &\quad - \frac{\Delta}{S} E[(1-\Delta) B | \bar{\mathbf{V}}(T), Y] \\ &= m(\bar{\mathbf{V}}(T), Y) + \int_0^T g(t, \bar{\mathbf{V}}(T), Y) \, dM(t), \qquad \text{(B.1)} \end{aligned}$$

where $A = a(\bar{\mathbf{V}}(T), Y), B = b(\bar{\mathbf{V}}(Q), Q), m(\bar{\mathbf{V}}(T), Y) = E[\Delta A + (1-\Delta) B | \bar{\mathbf{V}}(T), Y]$, and

$$g(t, \bar{\mathbf{V}}(T), Y) = b(\bar{\mathbf{V}}(t), t) + \frac{\int_0^t b(\bar{\mathbf{V}}(u), u) f_Q(u | \bar{\mathbf{V}}(T), Y) \, du}{S(t)} - \frac{m(\bar{\mathbf{V}}(T), Y)}{S(t)}$$

with $f_Q(\cdot | \bar{\mathbf{V}}(T), Y)$ the conditional density of Q given $(\bar{\mathbf{V}}(T), Y)$. The third identity in (B.1) follows from lemma 4.1 of Robins et al. (1999). Note that because A is arbitrary, so is $m(\bar{\mathbf{V}}(T), Y)$.

From the representation of $\Lambda(F_{R|L})^\perp$ just given, we conclude that $c(R, L_{(R)}) \in \Lambda(F_{R|L})^\perp$ if and only if $E[m(\bar{\mathbf{V}}(T), Y)] = 0$ and for all $t \in [0, T)$,

$$E\left[b(\bar{\mathbf{V}}(t), t) + \frac{\int_0^t b(\bar{\mathbf{V}}(u), u) f_Q(u | \bar{\mathbf{V}}(T), Y) \, du}{S(t)} - \frac{m(\bar{\mathbf{V}}(T), Y)}{S(t)} \,\middle|\, Q = t, \bar{\mathbf{V}}(t) \right] = 0. \qquad \text{(B.2)}$$

Equation (B.2) can be rewritten as the Volterra integral equation

$$b(\bar{\mathbf{V}}(t), t) = J_m(t) - \int_0^t b(\bar{\mathbf{V}}(u), u) f(u, t, \bar{\mathbf{V}}(t)) \, du, \qquad \text{(B.3a)}$$

where

$$J_m(t) = \frac{E[m(\bar{\mathbf{V}}(T), Y) \exp\{r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)\} | \bar{\mathbf{V}}(t)]}{E[S(t) \exp\{r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)\} | \bar{\mathbf{V}}(t)]} \qquad \text{(B.3b)}$$

and

$$f(u, t, \bar{\mathbf{V}}(t)) = \frac{E[\lambda_0(u | \bar{\mathbf{V}}(u)) S(u) \exp\{r(t, \alpha_0; \bar{\mathbf{V}}(T), Y) + r(u, \alpha_0; \bar{\mathbf{V}}(T), Y)\} | \bar{\mathbf{V}}(t)]}{E[S(t) \exp\{r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)\} | \bar{\mathbf{V}}(t)]}. \qquad \text{(B.3c)}$$

For each $m(\bar{\mathbf{V}}(T), Y)$, (B.3a) has a unique solution. Because $\Lambda_2^{O,\perp}$ is comprised precisely by all the functions of the observed data that belong to $\Lambda(F_{R|L})^\perp$, we conclude that

$$\begin{aligned} \Lambda_2^{O,\perp} = \Big\{ & \frac{\Delta}{S} m(\bar{\mathbf{V}}(T), Y) + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q) \\ & - \frac{\Delta}{S} E[(1-\Delta) b(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y] \colon \\ & E[m(\bar{\mathbf{V}}(T), Y)] = 0, b(\bar{\mathbf{V}}(t), t) \\ & \text{solves (B.3a) for } \forall t \in [0, T) \Big\}. \end{aligned}$$

Finally, intersecting $\Lambda_1^{O,\perp}$ and $\Lambda_2^{O,\perp}$, we obtain

$$\begin{aligned} \Lambda^{O,\perp} = \Big\{ & kH \colon \ k \in R^1 \text{ and } H \equiv \frac{\Delta}{S} \varepsilon + (1-\Delta) b(\bar{\mathbf{V}}(Q), Q) \\ & - \frac{\Delta}{S} E[(1-\Delta) b(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y], \\ & \text{where } b(\bar{\mathbf{V}}(t), t) \text{ solves (B.3a) with } m(\bar{\mathbf{V}}(T), Y) = \varepsilon \Big\}. \end{aligned} \qquad \text{(B.4)}$$

Because this set comprises multiples of the unique random variable H, we conclude that there is a single influence function for the parameter μ_0. This must be the case, because by Theorem 1, model $A(\alpha_0)$ is a nonparametric model for the observed data.

We now prove that $h(O; \mu_0, \Lambda_0; b^*)$ with b^* as in (7) is the unique $h(O, \mu_0, \Lambda_0; b)$ uncorrelated with all nuisance scores S_η. In the special case in which $\bar{\mathbf{V}}(t) = \mathbf{V}$ and $r(t, \alpha_0; \mathbf{V}, Y) = r(\alpha_0; \mathbf{V}, Y)$ for all t, (B.3a) has an explicit solution $b(\mathbf{V}, t) = b^*(\mathbf{V})$ independent of t given by

$$b^*(\mathbf{V}) = \frac{E[m(\mathbf{V}, Y) \exp\{r(\alpha_0; \mathbf{V}, Y)\} | \mathbf{V}]}{E[\exp\{r(\alpha_0; \mathbf{V}, Y)\} | \mathbf{V}]}.$$

In particular, when $m(\mathbf{V}, Y) = Y - \mu_0, b^*(\mathbf{V})$ coincides with $b^*(\mathbf{V}; \mu_0)$ in (7) and H in the set (B.4) coincides with $h(O, \mu_0, \Lambda_0; b^*)$. Thus $h(O, \mu_0, \Lambda_0; b^*)$ is uncorrelated with the nuisance scores S_η, because it is an element of $\Lambda_2^{O,\perp}$. The uniqueness of H follows from the uniqueness of b^*.

B.1.2 Computation of $\Lambda^{O,\perp}$ in Model $B(\alpha_0)$. It follows from the results of Ritov and Wellner (1988) that in model $B(\alpha_0)$,

$$\Lambda(F_{R|L}) = \left\{ \int_0^T \mathbf{g}(t)\, dM(t)\text{: } \mathbf{g}(t) \text{ is an arbitrary } (q+1) \text{ dimensional function of } t \right\}$$

and

$$\Lambda(F_{R|L})^{\perp} = \left\{ \mathbf{a}(\bar{\mathbf{V}}(T), Y) + \int_0^T \mathbf{g}(t, \bar{\mathbf{V}}(T), Y)\, dM(t)\text{: } E[\mathbf{g}(t, \bar{\mathbf{V}}(T), Y) | Q = t] = 0,\ E[\mathbf{a}(\bar{\mathbf{V}}(T), Y)] = 0, \text{ and } \mathbf{a} \text{ and } \mathbf{g} \text{ are } (q+1) \text{ dimensional} \right\}$$

We now use the representation (B.1) for any $\mathbf{c}(R, L_{(R)})$ and conclude that $\mathbf{c}(R, L_{(R)}) \in \Lambda(F_{R|L})^{\perp}$ if and only if $E[\mathbf{m}(\bar{\mathbf{V}}(T), Y)] = 0$ and for all $t \in [0, T)$,

$$E\left[\mathbf{b}(\bar{\mathbf{V}}(t), t) + \frac{\int_0^t \mathbf{b}(\bar{\mathbf{V}}(u), u) f_Q(u | \bar{\mathbf{V}}(T), Y)\, du}{S(t)} - \frac{\mathbf{m}(\bar{\mathbf{V}}(T), Y)}{S(t)} \,\middle|\, Q = t \right] = 0, \quad \text{(B.5)}$$

where $\mathbf{m}$ is now a $(q+1)$ dimensional function.

Equation (B.5) can be rewritten as

$$E[\mathbf{b}(\bar{\mathbf{V}}(t), t) D(t) + \mathbf{c}_{\mathbf{b},\mathbf{m}}(t)] = 0, \quad \text{(B.6)}$$

where

$$D(t) = S(t; \boldsymbol{\gamma}_0) \exp\{\boldsymbol{\gamma}_0' \mathbf{W}(t) + r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)\}$$

and

$$\begin{aligned} \mathbf{c}_{\mathbf{b},\mathbf{m}}(t) &= E\left[\exp\{\boldsymbol{\gamma}_0' \mathbf{W}(t) + r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)\} \times \int_0^t \mathbf{b}(\bar{\mathbf{V}}(u), u) f_Q(u | \bar{\mathbf{V}}(T), Y; \boldsymbol{\gamma}_0)\, du \right] \\ &\quad - E[\mathbf{m}(\bar{\mathbf{V}}(T), Y) \exp\{\boldsymbol{\gamma}_0' \mathbf{W}(t) + r(t, \alpha_0; \bar{\mathbf{V}}(T), Y)\}]. \end{aligned}$$

For any fixed function $\mathbf{m}(\bar{\mathbf{V}}(T), Y)$, (B.6) has an infinite number of solutions. But we now show that the set $\mathbf{b}^*_{\mathbf{m}}$ of solutions to (B.6) is equal to the set

$$\begin{aligned} \mathbf{b}^{**}_{\mathbf{m}} = \{ &\mathbf{b}(\bar{\mathbf{V}}(t), t) = \boldsymbol{\phi}(\bar{\mathbf{V}}(t), t) - E[D(t)]^{-1} \\ &\times \{E[\boldsymbol{\phi}(\bar{\mathbf{V}}(t), t) D(t)] + \mathbf{c}_{\mathbf{b},\mathbf{m}}(t)\}\text{:} \\ &\boldsymbol{\phi}(\bar{\mathbf{V}}(t), t) \text{ is an arbitrary } q+1 \text{ dimensional function}\}. \end{aligned}$$

That $\mathbf{b}^{**}_{\mathbf{m}} \subseteq \mathbf{b}^*_{\mathbf{m}}$ follows by direct verification that the elements of $\mathbf{b}^{**}_{\mathbf{m}}$ solve (B.6). That $\mathbf{b}^*_{\mathbf{m}} \subseteq \mathbf{b}^{**}_{\mathbf{m}}$ follows because for any arbitrary solution $\mathbf{b}(\bar{\mathbf{V}}(t), t)$ to (B.6),

$$\begin{aligned} &\mathbf{b}(\bar{\mathbf{V}}(t), t) \\ &\quad = \mathbf{b}(\bar{\mathbf{V}}(t), t) - E[D(t)]^{-1} \{E[\mathbf{b}(\bar{\mathbf{V}}(t), t) D(t)] + \mathbf{c}_{\mathbf{b},\mathbf{m}}(t)\}. \end{aligned}$$

Because $\Lambda_2^{O,\perp}$ is comprised precisely by all $(q+1)$-dimensional functions $\mathbf{c}(R, L_{(R)})$ of the observed data that belong to $\Lambda(F_{R|L})^{\perp}$, we conclude that

$$\begin{aligned} \Lambda_2^{O,\perp} = \Big\{ &\frac{\Delta}{S} \mathbf{m}(\bar{\mathbf{V}}(T), Y) + (1 - \Delta) \mathbf{b}(\bar{\mathbf{V}}(Q), Q) \\ &- \frac{\Delta}{S} E[(1 - \Delta) \mathbf{b}(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y]\text{:} \\ &E[\mathbf{m}(\bar{\mathbf{V}}(T), Y)] = 0 \text{ and } \mathbf{b}(\bar{\mathbf{V}}(t), t) \in \mathbf{b}^{**}_{\mathbf{m}} \Big\}. \end{aligned}$$

Finally, intersecting $\Lambda_1^{O,\perp}$ and $\Lambda_2^{O,\perp}$, we obtain

$$\begin{aligned} \Lambda^{O,\perp} = \Big\{ &\mathbf{H}\text{: } \mathbf{H} = \frac{\Delta}{S} \mathbf{k}\varepsilon + (1 - \Delta) \mathbf{b}(\bar{\mathbf{V}}(Q), Q) \\ &- \frac{\Delta}{S} E[(1 - \Delta) \mathbf{b}(\bar{\mathbf{V}}(Q), Q) | \bar{\mathbf{V}}(T), Y]\text{:} \\ &\mathbf{k} \in R^{q+1} \text{ and } \mathbf{b}(\bar{\mathbf{V}}(t), t) \in \mathbf{b}^{**}_{\mathbf{m}} \\ &\text{with } \mathbf{m}(\bar{\mathbf{V}}(T), Y) = \mathbf{k}\varepsilon \Big\}. \end{aligned}$$

This indeed shows that (10) characterizes the members of the set $\mathbf{b}^*$ of Section 4.2, because $\mathbf{b}^*(\bar{\mathbf{V}}(t), t; \boldsymbol{\psi}_0)$ satisfies (10) if and only if $\mathbf{b}^*(\bar{\mathbf{V}}(t), t; \boldsymbol{\psi}_0) \in \mathbf{b}^{**}_{\mathbf{m}}$ with $\mathbf{m}(\bar{\mathbf{V}}(T), Y) = k\mathbf{e}_1\varepsilon$ and $\mathbf{e}_1$ is the $(q+1)$ dimensional vector $(1, 0, \ldots, 0)'$. Further, it is simple to show that the set of influence functions corresponding to the class of estimators $\{\hat{\boldsymbol{\psi}}(\hat{\mathbf{b}}^*)\}$ contains all influence functions for $\boldsymbol{\psi}_0$.

APPENDIX C: AN ADAPTIVE ESTIMATOR

We propose an adaptive choice by $\hat{\boldsymbol{\phi}}_{\text{adap}}$ for the function $\boldsymbol{\phi}$. $\hat{\boldsymbol{\phi}}_{\text{adap}}$ will depend on both the data and the selection bias parameter α_0. The estimator $\hat{\mu}(\hat{\mathbf{b}}^*_{\text{adap}})$ determined by $\hat{\boldsymbol{\phi}}_{\text{adap}}$ via (11) will have asymptotic variance that should be close to the semiparametric variance bound for model $A(\alpha_0)$. To allow us to reuse the notation from Section 4.1, we assume that $r(t, \alpha_0; \bar{\mathbf{V}}(T), Y) = \alpha_0 Y$. We calculate $\hat{\boldsymbol{\phi}}_{\text{adap}}(\bar{\mathbf{V}}(t), t)$ using the following five-step procedure:

1. Obtain a preliminary *inefficient* RAL estimator $\hat{\boldsymbol{\psi}}(\mathbf{b}) = (\hat{\mu}(\mathbf{b}), \hat{\boldsymbol{\gamma}}(\mathbf{b})')'$ based on a convenient choice of $\mathbf{b}$ as in Section 4.1.
2. Specify a parametric model $f(\varepsilon, \bar{\mathbf{V}}(T); \boldsymbol{\eta})$ with $\boldsymbol{\eta}$ an unknown finite-dimensional parameter for the unknown law of $(\varepsilon, \bar{\mathbf{V}}(T))$ with $\varepsilon \equiv Y - \mu_0$.
3. Estimate $\boldsymbol{\eta}$ by the solution $\hat{\boldsymbol{\eta}}$ to the IPCW score equation
$$0 = \hat{E}_{\hat{\boldsymbol{\gamma}}(\mathbf{b})}[\partial \ln f(Y - \hat{\mu}(\mathbf{b}), \bar{\mathbf{V}}(T); \boldsymbol{\eta}) / \partial \boldsymbol{\eta}].$$
4. Replace the Volterra integral equation (B.3a)–(B.3c) by an estimated version where, in the estimated version, we replace
   a. in (B.3b) and (B.3c), $E[\cdot | \bar{\mathbf{V}}(t)]$ with expectations $E_{\hat{\eta}, \hat{\mu}(\mathbf{b})}[\cdot | \bar{\mathbf{V}}(t)]$ computed under $f(\hat{\varepsilon}(\mathbf{b}), \bar{\mathbf{V}}(T); \hat{\boldsymbol{\eta}})$ with $\hat{\varepsilon}(\mathbf{b}) \equiv Y - \hat{\mu}(\mathbf{b})$
   b. in (B.3b) and (B.3c), $S(\cdot)$ by $\hat{S}(\cdot | \bar{\mathbf{V}}(T), Y; \hat{\boldsymbol{\gamma}}(\mathbf{b}))$ of Section 4.1
   c. in (B.3c), $\lambda_0(u | \bar{\mathbf{V}}(u))$ by $\exp(\hat{\boldsymbol{\gamma}}(\mathbf{b})' \mathbf{W}(u)) d\hat{\Lambda}(u; \hat{\boldsymbol{\gamma}}(\mathbf{b}))$
   d. in (B.3c), $f(u, t, \bar{\mathbf{V}}(t))$ by $d\hat{F}(u, t, \bar{\mathbf{V}}(t))$
   e. in (B.3a), $f(u, t, \bar{\mathbf{V}}(t))\, du$ by $d\hat{F}(u, t, \bar{\mathbf{V}}(t))$.
5. Solve the estimated version of (B.3a)–(B.3c) and call the solution $\hat{\boldsymbol{\phi}}_{1,\text{adap}}(\bar{\mathbf{V}}(t), t)$, and define $\hat{\boldsymbol{\phi}}_{\text{adap}}(\bar{\mathbf{V}}(t), t)' = (\hat{\boldsymbol{\phi}}_{1,\text{adap}}(\bar{\mathbf{V}}(t), t), \mathbf{W}(t)')$.

Then, under mild regularity conditions, $\hat{\mu}(\hat{\mathbf{b}}^*_{\text{adap}})$ will be a RAL estimator under model $B(\alpha_0)$ with asymptotic variance equal to the semiparametric variance bound for model $A(\alpha_0)$ if the parametric model $f(\varepsilon, \bar{\mathbf{V}}(T); \boldsymbol{\eta})$ is correctly specified. Even under misspecification of the parametric model, $\hat{\mu}(\hat{\mathbf{b}}^*_{\text{adap}})$ will remain a RAL estimator under model $B(\alpha_0)$ with asymptotic variance that should remain close to the bound for model $A(\alpha_0)$ if η is a rather high-dimensional parameter.

[Received February 1998. Revised June 1999.]

REFERENCES

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Baker, S. G., Rosenberger, W. F., and DerSimonian, R. (1992), “Closed-Form Estimates for Missing Counts in Two-Way Contingency Tables,” *Statistics in Medicine*, 11, 643–657.

Bickel, P. J., Klaasen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

DeGruttola, V., and Tu, X. M. (1994), "Modeling the Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time," *Biometrics*, 50, 1003–1014.

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

Fitzmaurice, G. M., Clifford, P., and Heath, A. F. (1996), "Logistic Regression Models for Binary Panel Data With Attrition," *Journal of the Royal Statistical Society*, Ser. A, 159, 249–263.

Fitzmaurice, G. M., Laird, N. M., and Zahner, G. E. P. (1996), "Multivariate Logistic Models for Incomplete Binary Responses," *Journal of the American Statistical Association*, 91, 99–108.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. (1995), "Regression Models for Longitudinal Binary Responses With Informative Drop-Outs," *Journal of the Royal Statistical Society*, Ser. B, 57, 691–704.

Freedman, D., Rothenberg, T., and Sutch, R. (1984), "On Energy Policy Models," *Journal of Business and Economic Statistics*, 1, 24–36.

Good, I. J. (1983), *Good Thinking: The Foundations of Probability and Its Applications*, Minneapolis: University of Minnesota Press.

Gill, R. D. (1989), "Non- and Semi-Parametric Maximum Likelihood Estimators and the von Mises Method," *Scandinavian Journal of Statistics*, 16, 97–128.

Gill, R. D., and Robins, J. M. (1999), "Causal Inference from Complex Longitudinal Data: The Continuous Case," *The Annals of Statistics*, under review.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, 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. (1996), "A Trial Comparing Nucleoside Monotherapy With Combination Therapy in HIV-Infected Adults With CD4 Cell Counts from 200 to 500 per Cubic Millimeter," *The New England Journal of Medicine*, 15, 1081–1090.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244–2253.

Hogan, J. W., and Laird, N. M. (1997a), "Model-Based Approaches to Analyzing Incomplete Longitudinal and Failure Time Data," *Statistics in Medicine*, 16, 259–272.

——— (1997b), "Mixture Models for the Joint Distribution of Repeated Measures and Event Times," *Statistics in Medicine*, 16, 239–257.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Hougaard, P. (1986), "Survival Models for Heterogeneous Populations Derived From Stable Distributions," *Biometrika*, 73, 387–396.

Klein, J. P., and Moeschberger, M. L. (1988), "Bounds on Net Survival Probabilities for Dependent Competing Risks," *Biometrics*, 44, 529–538.

Laird, N. M. (1988), "Missing Data in Longitudinal Studies," *Statistics in Medicine*, 7, 305–315.

Lin, D. Y., and Ying, Z. (1994), "Semiparametric Analysis of the Additive Risk Model," *Biometrika*, 81, 61–71.

Little, R. J. (1985), "A Note About Models for Selectivity Bias," *Econometrica*, 53, 1469–1474.

——— (1993a), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.

——— (1993b), "Pattern-Mixture Models for Normal Missing Data," *Biometrika*, 81, 471–483.

——— (1995), "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R. J., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Mori, M., Woodworth, G. G., and Woolson, R. F. (1992), "Application of Empirical Bayes Inference to Estimation of Rate of Change in the Presence of Informative Right Censoring," *Statistics in Medicine*, 11, 621–631.

Murphy, S. (1995), "Likelihood Ratio–Based Confidence Intervals in Survival Analysis," *Journal of the American Statistical Association*, 90, 1399–1405.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Nordheim, E. V. (1984), "Inference From Nonrandomly Missing Categorical Data: An Example From a Genetic Study on Turner's Syndrome," *Journal of the American Statistical Association*, 7, 772–780.

Ritov, Y., and Wellner, J. A. (1988), "Censoring, Martingales, and the Cox Model," *Contemporary Mathematics*, 80, 191–219.

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

——— (1987), "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect (Addendum)," *Computers and Mathematics With Applications*, 14, 923–945.

——— (1997), "Non-Response Models for the Analysis of Non-Monotone Non-Ignorable Missing Data," *Statistics in Medicine*, 16, 21–38.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992), "Estimating Exposure Effects by Modeling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48, 479–495.

Robins, J. M., and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

Robins, J. M., and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology: Methodological Issues*, eds. N. Jewel, K. Dietz, and V. Farewell, Boston: Birkhauser.

——— (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Methods in Epidemiology*, ed. E. Halloran, New York: Springer-Verlag, 1–92.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

——— (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rotnitzky, A., and Robins, J. M. (1997), "Analysis of Semiparametric Regression Models With Non-Ignorable Non-Response," *Statistics in Medicine*, 16, 81–102.

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

Schluchter, M. D. (1992), "Methods for the Analysis of Informatively Censored Longitudinal Data," *Statistics in Medicine*, 11, 1861–1870.

Self, S., and Pawitan, Y. (1992), "Modeling a Marker of Disease Progression and Onset of Disease," in *AIDS Epidemiology: Methodological Issues*, eds. N. Jewel, K. Dietz, and V. Farewell, Boston: Birkhauser.

Slud, E., and Rubinstein, L. V. (1983), "Dependent Competing Risks and Summary Survival Curves," *Biometrika*, 70, 643–649.

Tricomi, F. G. (1957), *Integral Equations*, London: Interscience.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1994), "Modeling the Relationship of Survival and Longitudinal Data Measured With Error," *Journal of the American Statistical Association*, 90, 27–37.

van der Laan, M. (1993), "Efficient and Inefficient Estimation in Semiparametric Models," CWI Tract 114, Centre for Mathematics and Computer Science, Amsterdam.

Wu, M. C., and Bailey, K. R. (1988), "Analyzing Changes in the Presence of Informative Right Censoring Caused by Death and Withdrawal," *Statistics in Medicine*, 7, 337–346.

——— (1989), "Estimation and Comparison of Changes in the Presence of Informative Right Censoring: Conditional Linear Model," *Biometrics*, 45, 939–955. Corr.: 1990, 46, 88.

Wu, M. C., and Carroll, R. J. (1988), "Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process," *Biometrics*, 44, 175–188.

Zheng, M., and Klein, J. P. (1994), "A Self-Consistent Estimator of Marginal Survival Functions Based on Dependent Competing Risk Data and an Assumed Copula," *Communications in Statistics, Part A—Theory and Methods*, 23, 2299–2311.

——— (1995), "Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula," *Biometrika*, 82, 127–138.