Data-adaptive Selection Of The Adjustment Set In Variable Importance Estimation

Oliver Bembom[*] Jeffrey W. Fessel[†]
Robert W. Shafer[‡] Mark J. van der Laan[**]

[*]University of California, Berkeley, bembom@gmail.com

[†]Clinical Trials Unit, Kaiser Permanente, San Francisco, CA

[‡]Division of Infectious Diseases,Center for AIDS Research, Stanford University, Palo Alto, CA

[**]Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu

Data-adaptive Selection Of The Adjustment Set In Variable Importance Estimation

Oliver Bembom, Jeffrey W. Fessel, Robert W. Shafer, and Mark J. van der Laan

Abstract

If estimates of the effect of a treatment variable on an outcome of interest are to be adjusted for a set of possible confounding factors, it is necessary to rely on the assumption of experimental treatment assignment (ETA) according to which each experimental unit has positive probability of being observed at any of the possible levels of the treatment variable regardless of the values the confounding factors may take on. Even if this assumption is only practically violated in the sense that certain values of the confounding factors cause some treatment levels to become not impossible, but at least highly unlikely, the adjusted variable importance parameter often becomes poorly identified in finite samples.

We introduce an algorithm that is intended to make variable importance estimation more robust with respect to violations of the ETA assumption. Two different identifiability criteria are proposed for deciding when an adjusted variable importance parameter cannot be reliably estimated from the data. These criteria are then used to identify a maximal set of adjustment variables for which the ETA assumption appears reasonably well satisfied. A more efficient estimator of the parameter corresponding to the selected adjustment set is then sought by selecting from among estimators making use of even smaller adjustment sets by minimizing an estimate of the mean squared error for the selected parameter.

A simulation study aimed at evaluating the benefits of this latter step suggests that it can lead to efficiency gains on the order of 100% if the ETA assumption is violated to some extent and to efficiency gains on the order of 35% if the ETA assumption is well approximated. The proposed algorithm is applied to the problem of identifying mutations in the protease enzyme of HIV that have an effect on virologic response to the commonly used antiretroviral drug lopinavir. While both unadjusted and fully adjusted analyses yield unsatisfactory results, the subset of significant mutations identified by the algorithm introduced here includes eight of the 12 known major lopinavir resistance mutations as well as two mutations that are thought to increase susceptibility to lopinavir. Two of the four major mutations not identified in our analysis occurred very rarely in our data set, giving the algorithm low power to detect any impact on virologic response. Recent in vitro experiments suggest that the other two major mutations not identified here may in fact be less important in determining lopinavir resistance than previously thought. The excellent agreement of the results reported here with current understanding of lopinavir resistance suggest that variable importance estimation based on data-adaptive selection of the adjustment set represents a promising new approach for studying the effects of HIV mutations on clinical virologic response to antiretroviral therapy as well as for biomarker discovery in general.

1 Introduction

Many applications in modern biology measure a large number of genomic or proteomic covariates and are interested in assessing the impact of each of these covariates on a particular outcome of interest. In a study of HIV-positive patients, for example, a researcher may genotype the virus infecting each patient to ascertain the presence or absence of a large number of mutations, in the hope of identifying mutations that affect how a patient's plasma HIV RNA level (viral load) responds to a new drug regimen. Along with an estimate of the impact of each mutation on viral load, the researcher would generally like to have a measure of the statistical significance of these estimates in order to identify those mutations that are most likely to be genuinely related to the outcome. Such information could then be used to inform the decision of which drugs should be included in the regimen of a patient with a particular pattern of mutations.

The simplest way of assessing the impact of a particular mutation on viral load would be to compare the virologic response among patients whose virus has the mutation to that among patients whose virus does not. If we find that patients in the first group respond much more poorly to a particular drug regimen, a clinician might be inclined not to give this regimen to a new patient entering his office who has this mutation. Patients in the first group are, however, also quite likely to differ from those in the second group in terms of the remaining mutations as well as other measured clinical covariates. The mutation of interest may, for example, be very common among patients who have previously failed several similar drug regimens, making them far more likely to also fail the current one, but very rare among other patients. If the clinician's new patient comes from a population that differs from our original study population in that the mutation is not associated with having previously failed similar drug regimens, we might be wrong to conclude that the regimen under consideration would be a poor choice in this situation. Since the impact of the mutation of interest on viral load is confounded by the remaining mutations as well as other clinical covariates, such unadjusted estimates thus do not generalize to a new population in which the mutation of interest and the confounding factors are related to each other in a different way.

We might thus be interested in estimating the impact of a given mutation on viral load that is not due to associations of this mutation with any of the other measured covariates. Specifically, we might ask: What difference in virologic response would we observe if we could somehow give every patient in our study population the mutation interest, holding the remaining covariates fixed at their current values, as opposed to the scenario in which we give none of the patients this mutation, holding again other covariates fixed? Any observed difference could then not be due to differences of the two populations with regard to the remaining covariates and would thus be more likely to generalize to a new population in which the mutation of interest and the other covariates may be related to each other differently.

While such adjusted variable importance estimates are thus often more interesting than the corresponding unadjusted estimates, they also rely on an additional assumption in order to be identifiable from the collected data. Specifically, the assumption of experimental treatment assignment (ETA) requires that the adjustment variables cannot take on a set of values such that the group of patients corresponding to those values shows no variation in the mutation of interest. This assumption would be violated if, for example, there existed a second mutation that always occurred in concordance with the mutation of interest. Since we would never observe patients that exhibited each of the two mutations in the absence of the other one, it would be impossible to disentangle the individual effects of these two mutations, precluding us thus from estimating their impact on viral load adjusting for the other mutation.

More commonly, the set of adjustment variables may contain covariates that are not perfectly predictive of the mutation of interest, but that still determine the presence or absence of that mutation in a nearly deterministic fashion. A second mutation may, for example, be so strongly correlated with the mutation of interest that 99% of patients with this second mutation also exhibit the mutation of interest. In such instances, a substantial amount of data would be required before the adjusted variable importance of the mutation of interest could be estimated in any reliable way. In smaller samples, it could easily occur by chance that we observe no patients that are discordant for these two mutations, again precluding us from obtaining an adjusted variable importance estimate. To distinguish this scenario from the one described in the previous paragraph, we refer to it as a practical rather than a theoretical violation of the ETA assumption.

Under either of these two violations of the ETA assumption, the desired adjusted variable importance is not identifiable from the data at hand, making any estimates of this parameter unreliable and hard to interpret. A practical ETA violation, for example, often causes such estimates to become unstable and highly variable. An analysis that under such circumstances still aims to rank mutations based on adjusted variable importance estimates is thus bound to lead to unsatisfying results. Suppose, for example, that a mutation with no impact on viral load is strongly correlated with a second mutation that itself has a considerable impact. The practical ETA violation caused by this correlation would likely lead to highly variable and thus statistically non-significant adjusted variable importance estimates for both mutations. In this case, more useful results could be obtained by turning to variable importance estimates that do not attempt to adjust for the other mutation. This approach would likely yield significant estimates for both mutations, allowing the investigator to conclude that at least one of these two mutations has an impact on viral load. While we would have to acknowledge that we cannot disentangle the individual contributions of the two mutations, such a qualified identification of two mutations would generally seem preferable to the conclusion drawn from a fully adjusted analysis, according to which neither mutation would seem important in determining viral load.

These considerations suggest that it would be useful to have a criterion that could give the investigator a sense of the extent to which the variable importance parameter corresponding to a proposed adjustment set is identifiable from the data at hand. If this criterion suggested that the parameter corresponding to the full adjustment set was not well identified, it could then also be used to identify a smaller, more workable adjustment set. In this paper, we propose two criteria that can be used for this purpose, one based on the diagnostic for ETA bias developed by Wang et al. (2006), and one based on closed-form asymptotic bias estimates proposed by Bembom and van der Laan (2007a). In addition, we propose an approach for defining a sequence of nested candidate adjustment sets that, in combination with a given identifiability criterion, can be used to select an appropriate adjustment set data-adaptively.

Even if the variable importance parameter corresponding to a particular adjustment set is identified reasonably well by the data at hand, it may be advantageous to base estimation of this parameter on an adjustment set that in fact excludes additional covariates.The adjustment set defining the parameter of interest may, for example, contain a covariate that is a good predictor of the mutation under consideration, but only a weak predictor of viral load. Such a covariate will often be only a weak confounder of the relationship between the mutation and viral load, but can still lead to a mild practical violation of the ETA assumption that would cause the variable importance estimator to become more variable. Not adjusting for this covariate could thus, at the price of a slight increase in bias, offer a considerable reduction in variability, thus leading to an overall reduction in mean squared error. In this paper, we propose an approach that, given an adjustment set defining the parameter of interest, can be used to evaluate whether such additional exclusions from the adjustment set can be expected lead to more efficient estimates of that parameter. For the sake of clarity, we will refer to the adjustment set defining the parameter of interest, as selected based on a given identifiability criterion, as the targeted adjustment set; the possibly smaller adjustment set used in estimating this parameter, on the other hand, will be referred to as the effective adjustment set. The effective adjustment set is thus nested in the targeted adjustment set, which in turn is nested in the full adjustment set.

To summarize, this paper proposes an algorithm for variable importance estimation that first selects a targeted adjustment set defining the parameter of interest before then selecting an effective adjustment set that will be used in the estimation of this parameter. While the first step is aimed at making adjusted variable importance estimation robust to violations of the ETA assumption, the primary goal of the second step is to optimize efficiency. The remainder of the paper is organized as follows. In the next section, we review the formal definition of adjusted variable importance parameters as well as several estimators that have been proposed for these parameters. In section 3, we describe two different identifiability criteria that can be used for selecting the targeted adjustment set. The following section introduces our proposal for selecting the effective adjustment set. The possible efficiency gains that can be achieved by data-adaptively selecting the effective adjustment set are examined in a simulation study in section 5. Both steps of the proposed algorithm are then studied in an applied data analysis in section 6 that is aimed at ranking mutations in the protease enzyme of HIV based on their impact on virologic response to antiretroviral regimens containing the protease inhibitor lopinavir. We close with a brief discussion of possible extensions to the methodology described here.

2 Variable importance parameters and estimators

We consider the common point-treatment data structure consisting of n i.i.d. copies of $O = (W, A, Y)$, where $W = (W_1, \ldots, W_d)$ denotes the collection of measured confounders, A gives the treatment variable, and Y is the outcome of interest. For now we assume that A is binary. We would ideally like to estimate the marginal variable importance θ of A on Y controlling for W:

$$\theta \equiv E\Big[E[Y|A=1,W] - E[Y|A=0,W]\Big]. \tag{1}$$

This parameter is identified by the data under the ETA assumption according to which we have with probability 1.0 that, for $a \in \{0,1\}$,

$$P(A=a|W) > 0. \tag{2}$$

If there exists a w_1 such that $P(W=w_1) > 0$ and $P(A=a|W=w_1) = 0$ for $a=0$ or $a=1$, we say that the ETA assumption is theoretically violated. If (2) holds but there exists a w_2 such that $P(W=w_2) > 0$ and $P(A=a|W=w_2) \approx 0$ for $a=0$ or $a=1$, we say that the ETA assumption is practically violated.

We are here interested in identifying a maximal subset W^t of W such that we have with probability 1.0 that, for $a \in \{0,1\}$,

$$P(A=a|W^t) > \epsilon > 0, \tag{3}$$

assuring that the W^t-specific ETA assumption is neither theoretically nor practically violated. This in turn guarantees that the marginal variable importance of A on Y controlling for W^t can be identified from the data.

To identify the subset W^t, we first define a sequence of nested candidate adjustment sets. Since violations of the ETA assumption are caused by covariates that are highly predictive of A, we define these candidate adjustment sets based on a ranking of the confounders by their squared sample correlation with A. Specifically, each candidate adjustment set $W(\delta)$ will contain the δd covariates in W that have the lowest squared sample correlations with A, $0 \leq \delta \leq 1$. For this purpose, let $\rho_n^2(W_j, A)$ denote the squared sample correlation between W_j and A and let $q(\delta)$ denote the δ quantile of $\rho_n^2(W_1, A), \ldots, \rho_n^2(W_d, A)$. Then we can define $W(\delta) = (W_j : \rho_n^2(W_j, A) \leq q(\delta))$ as the collection of confounders with squared sample correlations no greater than the δ quantile $q(\delta)$ of squared sample correlations. The marginal variable importance parameter $\theta(\delta)$ corresponding to a candidate adjustment sets $W(\delta)$ is given by

$$\theta(\delta) \equiv E\Big[E[Y|A=1,W(\delta)] - E[Y|A=0,W(\delta)]\Big]. \tag{4}$$

Several estimators of such marginal variable importance parameters have been proposed (van der Laan, 2006). These estimators can typically be written as functions of the two nuisance parameters $g(\delta)(A,W) \equiv P(A|W(\delta))$ and $Q(\delta)(A,W) \equiv E[Y|A,W(\delta)]$. Assume that we have available preliminary estimates $g_n(\delta)$ and $Q_n(\delta)$ of these nuisance parameters; $g_n(\delta)$ may, for example, be obtained through a logistic regression of A on $W(\delta)$. Two popular variable importance estimators are then given by the G-computation estimator

$$\theta_n^{G-comp}(\delta) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta)(1,W_i) - Q_n(\delta)(0,W_i) \tag{5}$$

and the Inverse-Probability-of-Treatment-Weighted (IPTW) estimator

$$\theta_n^{IPTW}(\delta) = \frac{1}{n}\sum_{i=1}^{n} \Big[I(A_i=1) - I(A_i=0)\Big]\frac{Y_i}{g_n(\delta)(A_i,W_i)}. \tag{6}$$

The G-computation estimator yields valid estimates of $\theta(\delta)$ if $Q(\delta)$ is estimated consistently; the IPTW estimator instead relies on consistent estimation of $g(\delta)$.

Recently a targeted maximum-likelihood estimator of $\theta(\delta)$ has been proposed that gives consistent estimates as long as either $g(\delta)$ or $Q(\delta)$ is estimated consistently (van der Laan and Rubin, 2006). This estimator is identical to the G-computation estimator (5) except that it is based on an updated regression fit $Q_n^1(\delta)(A,W)$ rather than the initial fit $Q_n(\delta)(A,W)$. The updated estimate $Q_n^1(\delta)$ is obtained in a straightforward manner by adding the covariate

$$h(\delta)(A,W) = \left(\frac{I(A=1)}{g_n(\delta)(1,W)} - \frac{I(A=0)}{g_n(\delta)(0,W)}\right) \tag{7}$$

to the original regression fit and obtaining a maximum likelihood estimate $\epsilon_n(\delta)$ of the corresponding coefficient $\epsilon(\delta)$, holding all other coefficient estimates fixed at their initial values. The estimate $\epsilon_n(\delta)$ can thus be obtained by regressing Y on $h(\delta)(A,W)$ using $Q_n(A,W)$ as an offset. The updated regression fit $Q_n^1(\delta)(A,W)$ is then given by

$$Q_n^1(\delta)(A,W) = Q_n(\delta)(A,W) + \epsilon_n(\delta)h(\delta)(A,W). \tag{8}$$

The corresponding targeted maximum-likelihood estimate of $\theta(\delta)$ can be obtained as

$$\theta_n^{T-MLE}(\delta) = \frac{1}{n}\sum_{i=1}^{n} Q_n^1(\delta)(1,W_i) - Q_n^1(\delta)(0,W_i). \tag{9}$$

This estimator solves the estimating equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} D^{DR}(O_i|g_n(\delta), Q_n^1(\delta), \theta(\delta)) \tag{10}$$

corresponding to the double robust estimating function

$$\begin{aligned} D^{DR}(O|g(\delta),Q(\delta),\theta(\delta)) &= \Big[I(A=1) - I(A=0)\Big]\frac{Y - Q(\delta)(A,W)}{g(\delta)(A,W)} + \\ &\quad Q(\delta)(1,W) - Q(\delta)(0,W) - \theta(\delta). \end{aligned} \tag{11}$$

Under regularity conditions, the targeted maximum likelihood estimator is therefore asymptotically linear if at least one of the two nuisance parameters $g(\delta)$ and $Q(\delta)$ is estimated consistently (van der Laan and Robins, 2003). If both nuisance parameters are estimated consistently, the influence curve of the estimator is given by

$$IC^{T-MLE}(O|g_0(\delta), Q_0(\delta), \theta_0(\delta)) = c^{-1}D^{DR}(O|g_0(\delta), Q_0(\delta), \theta_0(\delta)), \tag{12}$$

where $g_0(\delta 0)$, $Q_0(\delta)$, and $\theta_0(\delta)$ denote the true values of the corresponding parameters and the standardizing constant c is given by

$$c = -\frac{\partial}{\partial\theta(\delta)} ED^{DR}(O|g_0(\delta), Q_0(\delta), \theta(\delta))\Big|_{\theta(\delta)=\theta_0(\delta)} = 1 \tag{13}$$

The asymptotic linearity of $\theta_n^{T-MLE}(\delta)$ under these conditions,

$$\sqrt{n}(\theta_n^{T-MLE}(\delta) - \theta_0(\delta)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC^{T-MLE}(O_i|g_0(\delta), Q_0(\delta), \theta_0(\delta)) + o_p(1), \tag{14}$$

implies in particular that

$$\sqrt{n}\Big(\theta_n^{T-MLE}(\delta) - \theta_0(\delta)\Big) \Rightarrow N\Big(0, \sigma^2(\delta) = Var(IC^{T-MLE}(O|g_0(\delta), Q_0(\delta), \theta_0(\delta))\Big). \tag{15}$$

An estimate $\sigma_n^2(\delta)$ of $\sigma^2(\delta)$ can be obtained as the sample variance of $IC^{T-MLE}(O|g_n(\delta), Q_n^1(\delta), \theta_n^{T-MLE}(\delta))$, allowing us to construct an asymptotic 95% confidence interval for $\theta(\delta)$ as

$$\theta_n^{T-MLE}(\delta) \pm 1.96\sqrt{\frac{\sigma_n^2(\delta)}{n}}. \tag{16}$$

If $g(\delta)$ is estimated consistently but $Q(\delta)$ is not, inference based on this approach is conservative (van der Laan and Robins, 2003).

Since the nuisance parameter $g(\delta)$ appears in the denominator of the covariate $h(\delta)(A, W)$, the targeted maximum-likelihood estimator can become quite unstable if some of the estimated treatment probabilities are close to zero, i.e. if the ETA assumption is practically violated. Its practical performance can often be improved somewhat by selecting a small value ϵ such as $\epsilon = 0.01$ and setting estimated treatment probabilities $g_n(\delta)(A, W) < \epsilon$ equal to ϵ.

We may also obtain a more stable estimator of $\theta(\delta)$ if we are willing to assume that the regression function $E[Y|A, W(\delta)]$ does not contain any interactions between A and $W(\delta)$. In that case, we have that

$$\theta(\delta) \equiv E\Big[E[Y|A = 1, W(\delta)] - E[Y|A = 0, W(\delta)]\Big] = E[Y|A = 1, W(\delta)] - E[Y|A = 0, W(\delta)]. \tag{17}$$

Under this additional modelling assumption, the targeted maximum-likelihood estimator of $\theta(\delta)$ no longer requires inverse weighting. Specifically, assume that we have available an estimate $g_n(\delta)$ and $Q_n(\delta)$ of the relevant nuisance parameters, with $Q_n(\delta)$ satisfying the additional modelling assumption so that it can be written as $Q_n(\delta) = \beta_n(\delta)A + r_n(W(\delta)) = \theta(\delta) + r_n(W(\delta))$ for some function $r_n(\cdot)$ of $W(\delta)$. As before, the targeted maximum-likelihood estimator is based on an updated estimate $Q_n^1(\delta)$ of $Q(\delta)$ that is obtained by adding a particular covariate $h(\delta)(A, W)$ to that initial fit. In this case, that covariate is given by

$$h(\delta)(A, W) = A - g_n(\delta)(1, W). \tag{18}$$

The updated fit for $Q(\delta)$ can then be written as $Q_n^1(\delta) = [\beta_n(\delta) + \epsilon_n(\delta)]A + r_n^1(W(\delta))$ so that the targeted maximum-likelihood estimator of $\theta(\delta)$ is given by

$$\theta_n^{T-MLE}(\delta) = \beta_n(\delta) + \epsilon_n(\delta). \tag{19}$$

In distinction to the non-parametric targeted maximum-likelihood estimator discussed previously, we will refer to this estimator as the model-based targeted maximum-likelihood estimator. This estimator solves the estimating function

$$0 = \frac{1}{n}\sum_{i=1}^{n} D(\delta)(O_i|g_n(\delta), r_n^1(\delta), \theta(\delta)) \tag{20}$$

corresponding to the estimating function

$$D(\delta)(O|g(\delta), r(\delta), \theta(\delta)) = \Big[A - g(\delta)(1, W)\Big]\Big[Y - \theta(\delta) - r(\delta)(A, W)\Big]. \tag{21}$$

Inference can thus again be based on the influence curve of this estimator. In this case, the standardizing constant c is given by

$$\begin{aligned} c &= -\frac{\partial}{\partial\theta(\delta)} ED(O|g_0(\delta), r_0(\delta), \theta(\delta))\Big|_{\theta(\delta)=\theta_0(\delta)} \\ &= A - g_0(\delta)(1, W). \end{aligned} \tag{22}$$

3 Selection of the targeted adjustment set

While the performance of all four estimators described above can be severely compromised if the ETA assumption is violated (Bembom and van der Laan, 2007b), the problems become most apparent in the case of the IPTW estimator. Unlike the other three estimators that under such circumstances can also rely on extrapolation through a correctly specified model for the regression $Q(\delta)$, the IPTW estimator is based entirely on inverse weighting by an estimate of the treatment mechanism $g(\delta)$, making it thus highly susceptible to violations of the ETA assumption. Under a theoretical violation, a subgroup of the target population is never observed at one of the possible treatment levels, preventing the re-weighting approach from successfully adjusting for confounding and thus resulting in biased estimates. Under a practical violation, observations with very small estimated treatment probabilities $g_n(\delta)$ and corresponding large weights tend to dominate the remainder of the sample so that the estimator becomes highly variable. In addition, it has been shown that a practical violation of the ETA assumption can in fact also cause the IPTW estimator to become biased in finite samples Neugebauer and van der Laan (2005). These considerations suggest that the finite-sample bias of this estimator is a useful measure of the degree to which a departure from the ETA assumption has caused the adjusted variable importance parameter $\theta(\delta)$ to become non-identifiable from the data at hand.

Wang et al. (2006) propose the following simulation-based approach for obtaining an estimate of this bias: The empirical distribution of $W(\delta)$ along with the nuisance parameter estimates $g_n(\delta)$ and $Q_n(\delta)$ define an estimate $P_n(\delta)$ of the data-generating distribution $P(\delta)$ for the observed data structure $O(\delta) = (W(\delta), A, Y)$. Under $P_n(\delta)$, the true value of the adjusted variable importance parameter $\theta(\delta)$ can be obtained by G-computation as

$$\theta(P_n(\delta)) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta)(1, W_i) - Q_n(\delta)(0, W_i). \tag{23}$$

At the same time, we can obtain a sampling distribution of IPTW estimates $\theta_{n,1}^{IPTW}(\delta), \ldots, \theta_{n,K}^{IPTW}(\delta)$ by applying the IPTW estimator to a large number K of realizations of the observed data structure $O(\delta)$ that were simulated under $P_n(\delta)$. The finite-sample bias of the IPTW estimator can then be estimated in a straightforward manner by

$$B_{sim}^{ETA}(\delta) = \frac{1}{K}\sum_{k=1}^{K} \theta_{n,k}^{IPTW}(\delta) - \theta(P_n(\delta)). \tag{24}$$

A possible limitation of this parametric bootstrap approach lies in its reliance on a large number of simulated data sets. First, such simulations can be computationally intensive so that the approach would not scale well to applications in which the group of candidate input variables for which we aim to obtain variable importance estimates is large. Second, unless an enormous number of data sets are simulated, the bias estimates can be expected to be quite sensitive to the exact number of simulated data sets used.

A computationally more tractable closed-form measure of finite-sample non-identifiability may be based on the following argument. A common recommendation for increasing the stability of IPTW estimators under practical ETA violations is to truncate the inverse-probability-of-treatment weights $1/g_n(\delta)(A, W)$ by some truncation constant M, thus using weights $wt_M = min(M, 1/g_n(\delta)(A, W))$ instead. Under a practical ETA violation, the use of such truncated weights can lead to a dramatic reduction in the variability of the IPTW estimator, but it typically also increases its bias. As long as at least some of the truncated weights wt_M are strictly less than the original weights $1/g_n(\delta)(A, W)$, the IPTW estimator will in fact often become asymptotically biased. Under a data-generating distribution that satisfies the ETA assumption, however, the estimated treatment probabilities $g_n(\delta)(A, W)$ are clearly bounded away from zero so that M would have to be chosen quite small for the truncated weights to become different from the original weights. Modest levels of truncation corresponding to a reasonably large value of M thus typically do not cause the IPTW estimator to become asymptotically biased if the ETA assumption is satisfied. These considerations suggest that the extent to which the ETA assumption is violated can also be quantified by the asymptotic bias of the IPTW estimator under modest truncation.

Bembom and van der Laan (2007a) recently derived a closed-form estimate for this bias as a function of the truncation constant M. Letting $g_{M,n}(\delta) \equiv max(g_n(\delta), M)$, this estimate is given by

$$\begin{aligned} B_M^{ETA}(\delta) &= \sum_{i=1}^{n} Q_n(\delta)(1, W_i)\frac{g_n(\delta)(1, W_i) - g_{M,n}(\delta)(1, W_i)}{g_{M,n}(\delta)(1, W_i)} - \\ & \sum_{i=1}^{n} Q_n(\delta)(0, W_i)\frac{g_n(\delta)(0, W_i) - g_{M,n}(\delta)(0, W_i)}{g_{M,n}(\delta)(0, W_i)} \end{aligned} \tag{25}$$

While this identifiability criterion is more computationally tractable than the simulation-based finite-sample bias estimate, it also requires the user to supply an appropriate truncation level M. The smaller M is chosen, the more sensitive $B_M^{ETA}(\delta)$ will be to practical ETA violations. At the same time, M should be chosen large enough to ensure that $B_M^{ETA}(\delta) = 0$ under a data-generating distribution that satisfies the ETA assumption. Since the IPTW estimator can tolerate larger weights as sample size increases, it would seem reasonable to make the selected truncation level a function of the sample size. One particular proposal would be to select M such that no observation in the sample would have a weight greater than some proportion p of the sum of all weights, where sensible choices for p corresponding to fairly modest levels of truncation might lie in the range from 0.05 to 0.20. We will examine the sensitivity of this proposed identifiability criterion to the exact choice of M in our data analysis in section 6.

It remains to define a reference with respect to which we evaluate the magnitude of the estimated bias $B_{sim}^{ETA}(\delta)$ or $B_M^{ETA}(\delta)$. A simple choice might be to consider the corresponding point estimate $\theta_n^{T-MLE}(\delta)$. Since the adjusted point estimates can become quite unreliable, however, if the ETA assumption is violated, we suggest to use the unadjusted variable importance estimate $\theta_n^{T-MLE}(0)$ instead. The targeted adjustment set $W^t = W(\delta^t)$ can now be selected by choosing δ^t to be the largest value of δ, $0 \leq \delta \leq 1$, such that the estimated bias is no greater than some proportion B_{max} of the unadjusted variable importance estimate. Here B_{max} is another user-supplied parameter, with reasonable choices likely to be made in the range from 0.10 to 0.50. We note that since the selection of $\theta(\delta^t)$ is made without knowledge of the point estimates $\theta_n^{T-MLE}(\delta)$, inference for $\theta(\delta^t)$ as based on the influence curve remains valid.

4 Selection of the effective adjustment set

Given a targeted adjustment set $W(\delta^t)$, we now aim to select an effective adjustment set $W^e = W(\delta^e)$ such that the corresponding estimator $\theta_n^{T-MLE}(\delta^e)$ has minimal mean squared error for estimating the targeted parameter $\theta(\delta^t)$. In many cases, the effective adjustment set will be equal to the targeted adjustment set, but if the targeted parameter still suffers from a mild practical ETA violation, it is possible that a smaller effective adjustment set will lead to a more efficient estimator.

The mean squared error for an estimator of $\theta(\delta^t)$ can be decomposed into the square of its bias and its variance. Since we will focus here on the non-parametric and model-based targeted maximum-likelihood estimators, the latter component can be estimated in a straightforward way based on the influence of these estimators (see section 2). We will use our estimates of the data-generating distributions $P_n(\delta)$ to obtain an estimate of the bias incurred by using a subset $W(\delta)$ of $W(\delta^t)$ in estimating $\theta(\delta^t)$. Under $P_n(\delta^t)$, the true parameter value of $\theta(\delta^t)$ is given by the G-computation estimate

$$\theta(P_n(\delta^t)) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta^t)(1, W_i) - Q_n(\delta^t)(0, W_i) = \theta_n^{G-comp}(\delta^t). \tag{26}$$

Under $P_n(\delta)$, $\delta \leq \delta^t$, the true parameter value of $\theta(\delta)$ is likewise given by

$$\theta(P_n(\delta)) = \frac{1}{n}\sum_{i=1}^{n} Q_n(\delta)(1, W_i) - Q_n(\delta)(0, W_i) = \theta_n^{G-comp}(\delta). \tag{27}$$

The desired bias can thus be estimated by the difference of the two relevant G-computation point estimates:

$$B_n^t(\delta) = \theta_n^{G-comp}(\delta) - \theta_n^{G-comp}(\delta^t) \tag{28}$$

We can now select δ^e as the minimizer over $0 \leq \delta \leq \delta^t$ of the corresponding mean squared error estimate

$$MSE_n^t(\delta) = \Big[B_n^t(\delta)\Big]^2 + V_n(\delta), \tag{29}$$

where $V_n(\delta)$ is an estimate of the variance of $\theta_n^{T-MLE}(\delta)$ as based on the influence curve of that estimator.

Since the selection of $\theta(\delta^e)$ is based on knowledge of the point estimates $\theta_n^{G-comp}(\delta)$, honest inference for the resulting estimator would have to take into account that it was selected from among several candidate estimators, specifically with the goal of minimizing mean squared error. Inference based on the influence curve of $\theta_n^{T-MLE}(\delta^e)$ may thus be somewhat optimistic since it ignores the data-adaptive selection of the estimator. Honest inference could be obtained based on a bootstrap procedure that includes this estimator selection step. Since we have that $B_n^t(\delta^t) = 0$, we note, however, that $\theta(\delta^e)$ can be expected to converge to $\theta(\delta^t)$ so that inference based on the influence curve of $\theta_n^{T-MLE}(\delta^e)$ remains asymptotically valid.

5 Simulation study

The selection of the targeted adjustment set is a question of selecting the scientific parameter of interest. The practical performance of the proposed approach to this problem is therefore better illustrated in an applied data analysis than in a simulation study. In this section, we present a simulation study that is aimed at examining to what extent the performance of the non-parametric and model-based targeted maximum-likelihood estimators of a given targeted variable importance parameter $\theta(\delta^t)$ can be improved by the data-adaptive selection of an effective adjustment set $W(\delta^e)$.

For this purpose, we consider a point-treatment data structure $O = (W, A, Y)$, with $W = (W_1, \ldots, W_{10})$ containing ten potential confounding factors, A denoting a binary treatment variable, and Y representing a continuous outcome of interest. Given a treatment mechanism $g_0(A \mid W)$ and the regression function $Q_0(A, W)$, the observed data structure was generated as follows:

1. Generate $W_1, \ldots, W_{10}$ as independent random uniform variables over the interval $[0, 1]$.
2. Generate the observed treatment variable A from the conditional distribution of A given W, $g_0(A \mid W)$.
3. Generate the observed outcome Y as $Y = Q_0(A, W) + \epsilon$ with $\epsilon \sim N(0, 1)$.

We consider the two different treatment mechanism

$$\text{logit}\Big(g_{1,0}(A \mid W)\Big) = W_3 - W_4 + 2W_5 - 2W_6 + 2W_7 - 2W_8 - 3W_9 + 4W_{10} \tag{30}$$

and

$$\text{logit}\Big(g_{2,0}(A \mid W)\Big) = W_3 - W_4 + 2W_5 - 2W_6 + 2W_7 - 2W_8 - 2W_9 + 2W_{10}. \tag{31}$$

The regression function $Q_0(A, W)$ is given by

$$Q_0(A, W) = A + 2W_2 + 2W_3 + 2W_4 + W_7 + W_8 + 0.1W_9 + 0.1W_{10}. \tag{32}$$

We thus have two different data-generating distributions $(g_{1,0}, Q_0)$ and $(g_{2,0}, Q_0)$. The targeted parameter is given by the fully adjusted marginal variable importance

$$\theta = E\Big[E[Y|A = 1, W] - E[Y|A = 0, W]\Big]. \tag{33}$$

Under $(g_{1,0}, Q_0)$, the covariates W_9 and W_{10} are strong predictors of A so that they may cause a moderate practical violation of the ETA assumption. Since they have only a weak effect on Y, omitting these two covariates from the effective adjustment set might therefore lead to a considerable reduction in the variability of the estimator, at the price of only a slight increase in bias. Data-adaptive selection of an effective adjustment set can therefore be hoped to lead to a significant increase in efficiency under this data-generating

Table 1: Mean squared error of the non-parametric and model-based targeted maximum-likelihood estimators using either the targeted adjustment set or a data-adaptively selected effective adjustment set.

	Non-parametric		**Model-based**	
	Targeted	Effective	Targeted	Effective
$(g_{1,0}, Q_0)$				
n = 100	16.7632	0.0917	0.0758	0.0670
n = 500	0.0312	0.0140	0.0138	0.0131
n = 2500	0.0051	0.0025	0.0027	0.0025
$(g_{2,0}, Q_0)$				
n = 100	1.9750	0.0828	0.0645	0.0621
n = 500	0.0168	0.0125	0.0119	0.0116
n = 2500	0.0030	0.0022	0.0022	0.0021

distribution. Under $(g_{2,0}, Q_0)$, W_9 and W_{10} are only moderate predictors of A so that much smaller efficiency gains might be expected under this data-generating distribution.

Table 1 summarizes the mean-squared errors for the non-parametric and model-based targeted maximum-likelihood estimators of θ using either the targeted adjustment set or a data-adaptively selected effective adjustment set for three different sample sizes. As expected, the fully adjusted non-parametric estimator is more sensitive to practical ETA violations than the fully adjusted model-based estimator, with its variance being considerably greater under $(g_{1,0}, Q_0)$ than under $(g_{2,0}, Q_0)$. Consequently, the non-parametric estimator also benefits much more strongly from the data-adaptive selection of an effective adjustment set, with efficiency gains relative to the fully adjusted estimator of roughly 100% under $(g_{1,0}, Q_0)$ and 35% under $(g_{2,0}, Q_0)$ for sample sizes of $n = 500$ and greater. The enormous efficiency gains observed for this estimator for $n = 100$ suggest a considerable practical ETA violation that in practice might have resulted in the selection of a smaller targeted adjustment set. The efficiency gains for the model-based estimators are slight compared to those for the non-parametric estimator. Since the assumption of no interaction between A and W is satisfied in this simulation study, the model-based estimator is typically more efficient than the non-parametric estimator. We note, however, that the performance of the two estimators based on a data-adaptively selected effective adjustment set is comparable, especially as sample size increases, which is another testament to the considerable efficiency gains achieved by the non-parametric estimator.

6 Data analysis

In this section we apply the methodology described above to the task of identifying mutations in the protease enzyme of HIV that modulate how well the virus can replicate in the presence of a particular antiretroviral drugs, and thus how well a patient responds to that drug. A considerable number of such drugs are available for treating patients infected with HIV, with the main mechanistic classes consisting of protease inhibitors (PIs), nucleotide and nucleoside reverse transcriptase inhibitors (NRTIs), and nonnucleoside reverse transcriptase inhibitors (NNRTIs). While a patient is being treated with a particular combination of these drugs, the virus frequently acquires a number of mutations that reduce its susceptibility to that drug regimen, requiring the patient to be switched to a new regimen that the virus remains sensitive to. When faced with this situation, clinicians frequently genotype the virus to ascertain the presence or absence of a large number of mutations that are thought to contribute to the resistance to various drugs (Shafer et al., 2000). This practice motivates us here to identify in a systematic way mutations that have a strong impact on a patient's virologic response to a new drug treatment and that could thus guide a clinician in designing a salvage therapy regimen on the basis of genotypic test results.

The effect of viral mutations on virologic response to therapy can be seriously confounded by a patient's treatment history. Past treatment regimens exert a strong selection pressure on viral evolution, thus affecting the probability that a given mutation is observed. In addition, treatment history can have an independent impact on virologic response by resulting in archived, or latent, virus carrying unobserved mutations that affect response to subsequent treatment regimens. As a result, an unadjusted association observed between a given mutation and treatment response may in fact be due to the presence of other mutations, both observed and unobserved. Treatment strategies vary across populations and evolve over time, potentially resulting in distinct mutation distributions. Thus, control of confounding due to treatment history is needed to ensure that the estimated importance of a given mutation can be more readily generalized to populations other than the original study population.

Similarly, we would also like to adjust for the presence of additional mutations. Mutations conferring resistance to drugs of a class different from that targeted by the mutation of interest, thus affecting a distinct viral enzyme, can typically be controlled for without much difficulty. However, mutations conferring resistance to the same drug class, thus affecting the same viral enzyme, are often so strongly correlated that the corresponding adjusted variable importance parameter is subject to a severe ETA violation. This is due to the fact that, while correlation between mutations affecting distinct viral enzymes occurs primarily as a result of past treatment patterns, correlation between mutations in the same enzyme often occurs as part of an evolutionary pathway towards resistance to drugs targeting that enzyme. Previous analyses have typically addressed this problem by categorically not adjusting for any mutations affecting the same viral enzyme as the mutation under consideration (Bembom et al., 2007). One might expect, however, that only a subset of the mutations affecting the same viral enzyme are so strongly correlated with the mutation under consideration as to cause serious ETA problems so that data-adaptive selection of the adjustment set might lead to variable importance estimates that typically suffer from less confounding than those obtained in earlier analyses.

6.1 Data source

Analyses were based on a data set, described previously in Bembom et al. (2007), consisting of observational clinical data that were primarily drawn from the Stanford drug resistance database and supplemented with data from an ongoing collaboration with the Kaiser Permanente Medical Care Program, Northern California. Currently, the Stanford database contains longitudinal data on over 6,000 patients. Data collected include use of antiretroviral drugs, results of viral genotype tests, and measurements of viral load as well as CD4 T cell count collected during the course of clinical care.

For the sake of illustration, we focus on resistance to the commonly used PI drug lopinavir. We identified all Treatment Change Episodes (TCEs) in this database that involved initiation of a salvage regimen containing lopinavir. A TCE was defined using the following inclusion criteria: 1) change of at least one drug from the patient's previous antiretroviral regimen; 2) availability of a baseline viral load and genotype within 24 weeks prior to the change in regimen; and, 3) availability of an outcome viral load 4-36 weeks after the change in regimen and prior to any subsequent changes in regimen.

TCEs were excluded if no candidate resistance mutations were present in the baseline genotype, if the subject had no past experience of PI drugs prior to the current regimen, or if the newly initiated regimen included hydroxyurea, any experimental antiretroviral drugs, or any PI drugs other than lopinavir (apart from the low dose of ritonavir that is always given with lopinavir). If a single baseline genotype had several subsequent regimen changes that met inclusion criteria as TCEs, only the first of these regimen changes was included in analyses. Multiple TCEs, each corresponding to a unique baseline genotype, treatment changes, and outcome, were allowed from a single individual; the resulting dependence between TCEs was accounted for in the derivation of standard errors and p-values. Based on these inclusion criteria, we identified 401 TCEs among 372 subjects that were included in our analyses. We considered as candidate biomarkers all mutations assessed by the Stanford HIVdb algorithm to be potentially related to resistance to any approved PI drug (`http://hivdb.stanford.edu`, accessed 9/1/2007). Including only mutations that occurred in at least two TCEs, we are faced with a total of 26 candidate PI mutations.

Antiretroviral regimens generally combine drugs from more than one class. The following characteristics of the non-PI component of the salvage regimen were therefore included in the set W of potential adjustment variables: indicators of use of each of 13 non-PI drugs; number of drugs used in each major non-PI class;

number of drugs and number of classes used in the salvage regimen for the first time; use of an NNRTI drug in the salvage regimen for the first time; and number of drugs switched between the previous and salvage regimen. W also included the following covariates collected prior to the baseline genotype: indicators of past treatment with each of 30 antiretroviral drugs; number of drugs used in each of the three major drug classes (PI, NRTI, and NNRTI); history of mono or dual therapy; number of past drug regimens; date of earliest antiretroviral therapy; highest prior viral load; lowest prior CD4 T cell count; and most recent (baseline) viral load.

The covariate set W also included indicators for the presence or absence of PI mutations other than the mutation of interest itself as well as indicators for the presence or absence of known NRTI and NNRTI mutations. In addition, we included summaries of the non-PI mutations in the baseline genotype. Known NRTI and NNRTI resistance mutations present at baseline were summed. Furthermore, susceptibility scores (standardized to a 0-1 scale) were calculated for each non-PI antiretroviral drug using the Stanford HIVdb scoring system. These susceptibility scores were included both as individual covariates and as interactions with indicators of the use of their corresponding drugs in the salvage regimen. Finally, these interaction terms were summed to yield a non-PI genotypic susceptibility score (GSS), which summarized the activity of the non-PI component of the regimen. The set of potential adjustment variables W included a total of 163 variables.

The outcome of interest, clinical virologic response, could be conceived as either a binary indicator of success (defined as achievement of a final viral load below the assay's lower limit of detection of 50 copies/mL), or as a continuous measure such as the change in final log_{10} viral load over baseline log_{10} viral load. The analyses reported here used a hybrid of these two approaches, aiming to capture the strengths of each. Specifically, given a baseline measurement Y_0 and a follow-up measurement Y_1 of log_{10} viral load, the outcome of interest Y was defined as follows: If Y_1 was above the lower limit of detection ($Y_1 > 1.7$), then $Y = Y_1 - Y_0$; if Y_1 was below the detectability limit, however, we imputed Y as the maximum decrease in viral load detected in the population, which was -4.2 log. Under this definition, both large drops in viral load from a high baseline and any achievement of an undetectable viral load (regardless of baseline) were treated as clinical successes. When several viral loads were measured between 4 and 36 weeks after regimen change, the first was used; duration from initiation of the salvage regimen until outcome measurement was included in the adjustment set W.

6.2 Variable importance estimation

The goal of our analysis was to estimate the impact of each of the 26 candidate PI mutations on Y, adjusting for as many elements of W as possible, and to rank the mutations based the statistical evidence for a non-zero variable importance. For this purpose we focus on the non-parametric and model-based targeted maximum-likelihood estimators described in section 2. We compare the results based on data-adaptively selected targeted and effective adjustment sets to those based on unadjusted and fully adjusted analyses.

Covariates that are not predictive of the outcome of interest neither confound the effect of a mutation on viral nor have the potential to increase the precision with which that effect can be estimated. Hence we first carried out a dimension reduction step aimed at identifying those covariates in W that appear to be associated with viral load. For this purpose, we examined the univariate association between each baseline covariate W_j and Y using a univariate repeated-measures regression. In this manner, we obtained p-values for the null hypotheses that a given covariate is independent of Y. Since the collection of candidate baseline covariates was sizeable, these marginal p-values were adjusted for the simultaneous performance of multiple hypothesis tests using the approach developed by Benjamini and Hochberg (1995) for controlling the false discovery rate (FDR). Out of the 163 variables contained in W, we retained a total of 51 that remained significantly associated with Y at a significance level of 0.05.

Following this dimension reduction, we applied the Deletion/Substitution/Addition (D/S/A) algorithm (Sinisi and van der Laan, 2004) to obtain estimates of the two nuisance parameters $g(\delta)$ and $Q(\delta)$. The D/S/A algorithm is a data-adaptive algorithm for polynomial regression that generates candidate predictors as linear combinations of polynomial tensor products in the candidate explanatory variables. These candidate estimators are indexed by the number and complexity of the terms, and the optimal candidate is selected using cross-validation. A version of the D/S/A algorithm was employed that relied solely on addition moves to generate candidate estimators, thus making it similar to a forward regression approach except that the size of the estimator is selected by cross-validation rather than by p-values; deletion and substitution moves were omitted to reduce computational complexity. The algorithm considered candidate estimators consisting of up to 20 terms.

Given a set of candidate explanatory variables, two-way interactions were explored based on repeated-measures regression models aimed at predicting Y as function of two candidate explanatory variables as well as the corresponding interaction term. Two-way interaction terms that were significant at an FDR-adjusted significance level of 0.05 were explicitly included in the set of candidate explanatory variables. The D/S/A algorithm was then allowed to consider estimators consisting only of main-effect terms taken from that set of candidate explanatory variables. This approach of not considering candidate estimators involving arbitrary two-way interactions is motivated not only by computational considerations, but also by the observation that such estimators are typically far more variable than those based on main-effect terms only, thus often leading to the selection of estimators including only main-effect terms. Including important interaction terms explicitly in the set of explanatory variables can thus be hoped to alleviate the discontinuity in variability seen in moving from estimators consisting of only main-effect terms to those involving also two-way interactions, thus increasing the chance that important two-way interaction terms will be selected in the final estimator.

Since we are interested in estimating the effect of a given mutation A on Y, we would like A to be included in the regression model for $E[Y|A,W]$. Forcing A into the model and then allowing the D/S/A algorithm to add elements of W is problematic, however, since adjustment variables that are strongly correlated with A may contribute little to the accurate prediction of Y once A is included in the model. Such an approach might thus lead to important confounding factors being omitted from the model. We therefore first allow the D/S/A algorithm to data-adaptively select a linear regression model for $E[Y|W]$ before then re-fitting that model with A added to the selected explanatory variables.

The D/S/A algorithm was also used to select an appropriate logistic regression model for the treatment mechanism $g(\delta)$. The selection criterion of minimizing cross-validated risk employed by the D/S/A algorithm for selecting the size of the estimator is aimed at selecting an estimator with good prediction properties. Optimizing the bias-variance trade-off for this purpose often leads to estimators consisting of only a small number of terms, causing the selected regression fit for $g(\delta)$ to give an unrealistically optimistic impression of the extent to which the ETA assumption is satisfied. For this reason, it is typically advisable to use somewhat more non-parametric estimates of the treatment mechanism for the task of assessing the validity of the ETA assumption (Wang et al., 2006). We therefore selected the size of the regression fit for g not as the minimizer of cross-validated risk, but rather as the largest size such that the corresponding cross-validated risk was no more than 25% above the minimal cross-validated risk. Theoretical arguments in fact imply that such slight overfits of the treatment mechanism will in first order also increase the efficiency of the resulting variable importance estimator (van der Laan and Robins, 2003). Overfits may in theory negatively affect the performance of the estimator through second-order terms if some of the estimated treatment probabilities become very close to zero, but the variable importance algorithm proposed here addresses that problem by selecting a targeted adjustment set for which the ETA assumption is well approximated. In addition, we set estimated treatment probabilities smaller than 0.01 to 0.01.

Point estimates based on the approach presented in section 2 for a sample of n i.i.d. observations remain valid in the context of repeated measures. The efficiency of the estimator might be improved by optimizing the weights given to individual observations on basis of an estimated correlation matrix for observations obtained from the same subject, as is done in generalized estimating equations Liang and Zeger (1986), but given the small number of repeated measures in the data set at hand we simply give equal weights to all observations. Estimation is thus based on estimating functions that are a sum over all the observations contributed by a single subject. Assuming that the number of observations contributed by each subject is non-informative, inference can be based in a straightforward manner on these modified estimating functions.

6.3 Unadjusted and fully adjusted variable importance estimates

Among the 26 candidate PI mutations considered here, the Stanford scoring system identifies the following 12 mutations as major contributors to lopinavir resistance: 50V, 82AFST, 46ILV, 54VA, 54LMST, 84AV, 32I, 47V, 48VM, 82MLC, 84C, and 90M; the remaining 14 mutations are thought to make minor or no contributions to resistance. Here and subsequently, mutations are denoted by the position of the change in the HIV protease enzyme, followed by a letter indicating the amino acid that has been substituted (e.g. 53LY refers to a substitution of leucine or tyrosine at protease position 53).

The unadjusted variable importance analysis, summarized in table 2, yielded significant p-values for all but eight of the candidate PI resistance mutations. Four of these eight mutations occurred in fewer than 10 TCEs so that the analysis had low power to detect an impact of these mutations on viral load. Among these four mutations were two mutations, 82MLC and 84C, that are thought to have a major effect on lopinavir resistance. The significant subset includes the remaining 10 known major lopinavir resistance mutations, but also eight mutations thought to make minor or no contributions to resistance. Among these were the mutations 30N, 88DTG, and 88S, all estimated to be significantly protective. Under the Stanford scoring system, mutations only receive a score if they are thought to increase resistance to a particular drug so that these findings are not necessarily in disagreement with the scores of zero assigned to these three mutations by that system. It seems quite plausible that mutations may also decrease the fitness of the virus and thus lead to improved virologic response. In fact, *in vitro* experiments examining the effect of different mutations on viral phenotype suggest that 30N and 88S may in fact have a negative impact on the fitness of the virus (Rhee et al., 2006). The significant subset still contains five mutations, however, that are estimated to be associated with considerably worse virologic response, but are not considered major lopinavir drug resistance mutations by the Stanford scoring system (33F, 73CSTA, 10FIRVY, 20IMRTVL, and 71ITV). Two of these, 33F and 73CSTA, are in fact ranked among the five most important mutations by the unadjusted analysis, illustrating that an analysis not addressing the issue of confounding can lead to rather noisy results.

An analysis based on the non-parametric targeted maximum-likelihood estimator adjusting for the full set W of potential confounders, summarized in table 3, identifies only five mutations as having a major effect on virologic response to lopinavir (50V, 84C, 16E, 32I, and 48VM). In agreement with the Stanford scoring system, the two mutations 50V and 32I are estimated to lead to decreased susceptibility to lopinavir. The mutation 16E, estimated to lead to considerably improved fitness of the virus in the presence of lopinavir, is not thought to be a major contributor to lopinavir drug resistance. The remaining two mutations 84C and 48VM, finally, are thought to be major contributors, but are in fact estimated to lead to increased susceptibility to lopinavir. The ranking produced by this analysis is thus hard to reconcile with current understanding of HIV antiretroviral resistance, illustrating that a fully adjusted analysis can lead to unreliable results if the ETA assumption is violated. A fully adjusted analysis based on the model-based targeted maximum-likelihood estimator, summarized in table 4, identified no mutations with a statistically significant impact on virologic response. These findings are similarly unsatisfying and show that a violation of the ETA assumption cannot be adequately addressed by turning to a more stable estimator that is based on additional modelling assumptions.

6.4 Data-adaptive selection of the targeted and effective adjustment sets

In this section, we examine the variable importance estimates obtained by data-adaptive selection of the targeted and effective adjustment set. In section 3, we proposed two different criteria for selecting the targeted adjustment set, one based on a simulation aimed at estimating the finite-sample bias of the IPTW estimator, the other based on a closed-form estimate of the asymptotic bias of a modestly truncated IPTW estimator. The latter criterion depends on a user-supplied choice for the parameter p that defines the desired truncation level based on the maximum proportion of the sum of all weights that any one weight is allowed to reach. We first examine the sensitivity of the proposed algorithm to different choices for selecting the targeted adjustment set. Table 5 summarizes the targeted adjustment level δ^t selected by the simulation-based criterion as well as by the closed-form criterion for three different choices of p. Overall, the choices made by the four different approaches are in good agreement with each other, with major discrepancies

Table 2: Unadjusted estimates variable importance estimates ranked by p-value. *The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	p-value
1	30N	0	45	-1.12	0.23	0.0000
2	54VA	11	84	0.86	0.18	0.0000
3	50V	20	5	1.98	0.44	0.0001
4	33F	5	44	0.84	0.21	0.0005
5	73CSTA	2	66	0.81	0.22	0.0012
6	88DTG	0	44	-0.89	0.25	0.0012
7	82AFST	20	100	0.62	0.18	0.0016
8	10FIRVY	2	217	0.54	0.16	0.0022
9	90M	10	171	0.54	0.16	0.0028
10	47V	10	17	1.18	0.36	0.0029
11	54LMST	11	36	0.69	0.24	0.0110
12	88S	0	9	-0.74	0.27	0.0134
13	32I	10	21	0.80	0.30	0.0146
14	20IMRTVL	0	115	0.46	0.18	0.0182
15	46ILV	11	143	0.43	0.17	0.0182
16	84AV	11	73	0.49	0.20	0.0219
17	71TVI	2	181	0.36	0.16	0.0312
18	48VM	10	16	0.77	0.35	0.0378
19	53LY	3	33	0.53	0.26	0.0601
20	24IF	2	16	0.69	0.36	0.0691
21	36ILVTA	0	141	0.32	0.18	0.0919
22	23I	0	4	0.68	1.02	0.5950
23	63P	0	311	0.09	0.19	0.7297
24	82MLC	10	4	0.30	0.95	0.8123
25	16E	0	9	-0.05	0.50	0.9308
26	84C	10	2	0.15	1.74	0.9308

Table 3: Fully adjusted non-parametric variable importance estimates ranked by p-value. *The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	p-value
1	50V	20	5	0.95	0.08	0.0000
2	84C	10	2	-2.92	0.08	0.0000
3	16E	0	9	0.90	0.25	0.0021
4	32I	10	21	0.26	0.07	0.0021
5	48VM	10	16	-0.26	0.07	0.0021
6	88S	0	9	-0.35	0.15	0.0708
7	33F	5	44	1.17	0.51	0.0823
8	54VA	11	84	0.55	0.28	0.1583
9	84AV	11	73	0.44	0.28	0.3457
10	53LY	3	33	0.51	0.38	0.4573
11	30N	0	45	-0.22	0.18	0.5307
12	47V	10	17	0.76	0.65	0.5307
13	10FIRVY	2	217	-0.22	0.21	0.6000
14	23I	0	4	-0.86	1.01	0.7144
15	73CSTA	2	66	0.20	0.25	0.7144
16	82AFST	20	100	-0.25	0.36	0.7824
17	82MLC	10	4	-0.33	0.61	0.9031
18	20IMRTVL	0	115	0.08	0.19	0.9098
19	36ILVTA	0	141	0.09	0.22	0.9098
20	90M	10	171	0.20	0.46	0.9098
21	63P	0	311	-0.09	0.29	0.9120
22	88DTG	0	44	-0.17	0.53	0.9120
23	46ILV	11	143	-0.04	0.21	0.9371
24	54LMST	11	36	-0.04	0.22	0.9371
25	71TVI	2	181	0.02	0.18	0.9651
26	24IF	2	16	0.00	0.27	0.9967

Table 4: Fully adjusted model-based variable importance estimates ranked by p-value. *The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	p-value
1	50V	20	5	1.35	0.54	0.2621
2	54VA	11	84	0.57	0.25	0.2621
3	16E	0	9	0.45	0.36	0.6008
4	24IF	2	16	0.54	0.31	0.6008
5	30N	0	45	-0.38	0.31	0.6008
6	33F	5	44	0.36	0.31	0.6008
7	36ILVTA	0	141	0.24	0.19	0.6008
8	47V	10	17	0.57	0.52	0.6008
9	48VM	10	16	-0.44	0.36	0.6008
10	53LY	3	33	0.30	0.26	0.6008
11	73CSTA	2	66	0.39	0.25	0.6008
12	88S	0	9	-0.47	0.33	0.6008
13	32I	10	21	0.37	0.36	0.6102
14	10FIRVY	2	217	-0.16	0.18	0.7016
15	20IMRTVL	0	115	0.08	0.16	0.8481
16	54LMST	11	36	0.17	0.30	0.8481
17	63P	0	311	-0.07	0.18	0.8481
18	71TVI	2	181	0.06	0.15	0.8481
19	84AV	11	73	0.11	0.22	0.8481
20	84C	10	2	-0.38	0.80	0.8481
21	88DTG	0	44	0.11	0.30	0.8481
22	90M	10	171	0.13	0.19	0.8481
23	23I	0	4	-0.23	1.22	0.9214
24	82MLC	10	4	-0.18	0.79	0.9214
25	46ILV	11	143	0.02	0.18	0.9275
26	82AFST	20	100	0.03	0.26	0.9275

observed only for the mutation 24IF. As is to be expected, larger choices for p, corresponding to milder truncation levels, decrease the sensitivity of the closed-form criterion and thus tend to lead to slightly larger targeted adjustment levels, although the effect is not too strong over the range of candidate values for p considered here.

Table 5: The targeted adjustment level δ^t selected based on the simulation-based criterion $B_{sim}^{ETA}(\delta)$ as well as based on the asymptotic criterion $B_M^{ETA}(\delta)$ for $p = 0.05$, $p = 0.10$, and $p = 0.20$. *The maximally tolerated proportion of bias relative to the unadjusted estimate, B_{max}, is set to 0.25.*

Mutation	Simulation	$p = 0.05$	$p = 0.10$	$p = 0.20$
10FIRVY	1.0	1.0	1.0	1.0
16E	0.0	0.0	0.0	0.0
20IMRTVL	1.0	1.0	1.0	1.0
23I	0.0	0.3	0.3	0.3
24IF	0.3	0.1	0.7	0.7
30N	0.5	0.5	0.5	0.5
32I	0.5	0.5	0.5	0.5
33F	0.8	0.7	0.8	0.9
36ILVTA	1.0	1.0	1.0	1.0
46ILV	1.0	1.0	1.0	1.0
47V	0.5	0.5	0.7	0.7
48VM	0.5	0.5	0.5	0.5
50V	1.0	1.0	1.0	1.0
53LY	0.9	0.8	0.9	0.9
54LMST	0.3	0.3	0.3	0.5
54VA	1.0	0.8	0.9	0.9
63P	0.4	0.6	0.6	0.6
71TVI	1.0	1.0	1.0	1.0
73CSTA	0.7	0.6	0.7	0.7
82AFST	0.8	0.7	0.7	0.8
82MLC	0.0	0.2	0.2	0.2
84AV	0.5	0.5	0.8	0.8
84C	0.0	0.0	0.0	0.0
88DTG	0.8	0.7	0.7	0.8
88S	0.0	0.2	0.4	0.4
90M	1.0	1.0	1.0	1.0

Table 6 summarizes the mutations that are identified as having a significant impact on virologic response if the targeted adjustment level is selected based on the same four different approaches. Again, the results are good agreement with each other, especially for those mutations that are thought to have major effect on virologic response. These findings, together with the results displayed in table 5, thus suggest that the proposed algorithm is fairly robust with respect to choices made at this step.

Table 7 summarizes the number of mutations that are statistically significant at the 0.05 level if the effective adjustment is either set equal to the targeted adjustment or selected data-adaptively based on the mean-squared-error criterion described in section 4. As seen already in the simulation study, the gains achieved by the non-parametric estimator are considerably larger than those achieved by the model-based estimator. The non-parametric estimator becomes more sensitive to the approach taken for selecting the targeted adjustment set if the effective adjustment set is not selected data-adaptively, with the number of significant mutations identified by that estimator ranging from three to eight depending on the choice of the identifiability criterion. The model-based estimator, on the other hand, remains relatively stable with

Table 6: Mutations that have a statistically significant impact on viral load at the 0.05 significance level. *Significant mutations are shown by check marks for the non-parametric (NP) as well as model-based (MOD) estimator and targeted adjustment levels* δ^t *selected based on the simulation-based criterion* $B_{sim}^{ETA}(\delta)$ *as well as based on the asymptotic criterion* $B_M^{ETA}(\delta)$ *for* $p = 0.05$, $p = 0.10$, *and* $p = 0.20$. *The maximally tolerated proportion of bias relative to the unadjusted estimate,* B_{max}, *is set to 0.25. The effective adjustment set is selected data-adaptively. The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Mutation	Score	Freq	Simulation	$p = 0.05$	$p = 0.10$	$p = 0.20$
10FIRVY	2	217				
16E	0	9				
20IMRTVL	0	115				
23I	0	4				
24IF	2	16	✓		✓	✓
30N	0	45	✓	✓	✓	✓
32I	10	21	✓	✓	✓	✓
33F	5	44				
36ILVTA	0	141				
46ILV	11	143				
47V	10	17	✓	✓	✓	✓
48VM	10	16	✓	✓	✓	✓
50V	20	5	✓	✓	✓	✓
53LY	3	33				
54LMST	11	36	✓	✓	✓	✓
54VA	11	84	✓	✓	✓	✓
63P	0	311				
71TVI	2	181				
73CSTA	2	66				
82AFST	20	100	✓	✓	✓	✓
82MLC	10	4				
84AV	11	73	✓	✓	✓	✓
84C	10	2				
88DTG	0	44				
88S	0	9	✓	✓	✓	✓
90M	10	171				

respect to that choice even if the effective adjustment set is not selected data-adaptively.

Table 7: Number of mutations that are statistically significant at the 0.05 significance level if the effective adjustment set is selected data-adaptively versus being set equal to the targeted adjustment set. *Results are shown for the non-parametric and model-based estimator as well as for the targeted adjustment set selected based on the simulation-based criterion $B_{sim}^{ETA}(\delta)$ as well as based on the asymptotic criterion $B_{M}^{ETA}(\delta)$ for $p = 0.05$, $p = 0.10$, and $p = 0.20$. The maximally tolerated proportion of bias relative to the unadjusted estimate, B_{max}, is set to 0.25.*

	Non-parametric		**Model-based**	
	Targeted	Effective	Targeted	Effective
Simulation	8	11	13	13
$p = 0.05$	8	10	11	12
$p = 0.10$	4	11	11	13
$p = 0.20$	3	11	10	11

Tables 8 and 9 summarize the variable importance estimates obtained by the algorithm proposed here, selecting the targeted adjustment by the closed-form criterion with $p = 0.05$. The non-parametric estimator identifies 10 mutations with a statistically significant impact on viral load. With the exception of 32I and 88S, all of these 10 mutations are also significant if the effective adjustment set is not selected data-adaptively. Among the 10 identified mutations are eight of the 12 major known drug resistance mutations for lopinavir (50V, 48VM, 47V, 54LMST, 32I, 54VA, 84AV, and 82AFST) as well as two mutations that are estimated to increase susceptibility to lopinavir (30N and 88S), a finding that, as mentioned earlier, is in agreement with *in vitro* experiments examining the effect of different mutations on viral phenotype (Rhee et al., 2006). The same experiments suggest that the mutations 46ILV and 90M, two of the four major mutations not identified by this analysis, may in fact be less important for lopinavir resistance than previously thought. The remaining two important mutations not identified here, 82MLC and 84C, occurred among only four and two of the 401 TCEs used in this analysis, respectively, so that the analysis had very low power for finding a significant impact of these mutations on viral load. Overall, the results reported here are thus in excellent agreement with current understanding of HIV antiretroviral resistance.

The variable importance estimates obtained by the model-based estimator are overall very similar to those obtained by the non-parametric estimator. The significant subset is identical except that the major mutation 84AV is missing and the three minor mutations 33F, 73CSTA, and 88DTG are included. With the exception of 88S, all of the identified 12 mutations are also significant if the effective adjustment set is not selected data-adaptively. The mutation 88DTG is estimated to increase susceptibility to lopinavir so that inclusion of this mutation is not necessarily in disagreement with the Stanford scoring system. The remaining three differences between the significant subset identified here and that described for the non-parametric estimator, however, cause the results for the model-based estimator to be in not quite as strong an agreement with current knowledge about lopinavir drug resistance as those for the non-parametric estimator.

For each of the mutations identified by the non-parametric estimator as a having a significant impact on viral load, table 10 summarizes which of the other significant mutations could not be adjusted for in obtaining an adjusted variable importance estimate. The table illustrates that adjustment for all other mutations is in fact difficult in most cases. Individual contributions to drug resistance are particularly hard to disentangle since mutations thought to decrease sensitivity to lopinavir are typically positively correlated with each other, but negatively with those mutations thought to increase sensitivity. For most candidate PI mutations it is still possible, however, to adjust for the majority of the other mutations. This may explain why the results reported here are in better agreement with current understanding of lopinavir resistance than those reported in previous analyses that categorically did not adjust for any other candidate PI mutations (Bembom et al., 2007). It seems somewhat surprising that even mutations with relatively small marginal correlations with the mutation of interest could sometimes not be adjusted for. Perhaps it is only when

Table 8: Data-adaptively adjusted non-parametric variable importance estimates ranked by p-value. *The targeted adjustment level is selected based on the asymptotic bias estimate $B_M^{ETA}(\delta)$ for a truncated IPTW estimator. The parameters p and B_{max} are set to 0.05 and 0.25, respectively. The effective adjustment set is selected data-adaptively. δ^t and δ^e give the proportion of potential confounders contained in the targeted and effective adjustment set, respectively. The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	δ^t	δ^e	p-value
1	50V	20	5	0.95	0.08	1.0	1.0	0.0000
2	48VM	10	16	1.20	0.18	0.5	0.5	0.0000
3	47V	10	17	1.62	0.25	0.5	0.4	0.0000
4	30N	0	45	-1.12	0.22	0.5	0.5	0.0000
5	54LMST	11	36	0.60	0.17	0.3	0.3	0.0025
6	32I	10	21	0.89	0.26	0.5	0.3	0.0027
7	88S	0	9	-0.74	0.27	0.2	0.0	0.0230
8	54VA	11	84	0.43	0.16	0.8	0.8	0.0251
9	84AV	11	73	0.44	0.17	0.5	0.5	0.0251
10	82AFST	20	100	0.38	0.15	0.7	0.6	0.0389
11	53LY	3	33	0.56	0.25	0.8	0.3	0.0521
12	73CSTA	2	66	0.54	0.24	0.6	0.5	0.0521
13	24IF	2	16	0.64	0.32	0.1	0.1	0.0947
14	33F	5	44	0.55	0.29	0.7	0.7	0.1068
15	36ILVTA	0	141	0.27	0.15	1.0	0.5	0.1333
16	90M	10	171	0.30	0.17	1.0	0.8	0.1333
17	88DTG	0	44	-0.32	0.29	0.7	0.7	0.4034
18	10FIRVY	2	217	-0.22	0.21	1.0	1.0	0.4333
19	20IMRTVL	0	115	0.13	0.15	1.0	0.9	0.5338
20	82MLC	10	4	0.47	0.58	0.2	0.2	0.5402
21	63P	0	311	-0.06	0.16	0.6	0.5	0.8915
22	23I	0	4	0.24	0.85	0.3	0.3	0.9133
23	16E	0	9	-0.05	0.50	0.0	0.0	0.9516
24	46ILV	11	143	0.01	0.18	1.0	0.9	0.9516
25	71TVI	2	181	0.02	0.18	1.0	1.0	0.9516
26	84C	10	2	0.15	0.87	0.0	0.0	0.9516

Table 9: Data-adaptively adjusted model-based variable importance estimates ranked by p-value. *The targeted adjustment level is selected based on the asymptotic bias estimate $B_M^{ETA}(\delta)$ for a truncated IPTW estimator. The parameters p and B_{max} are set to 0.05 and 0.25, respectively. The effective adjustment set is selected data-adaptively. δ^t and δ^e give the proportion of potential confounders contained in the targeted and effective adjustment set, respectively. The table also shows the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 9/1/2007) and the frequency of the mutation among the 401 treatment change episodes.*

Rank	Mutation	Score	Freq	Estimate	SE	δ^t	δ^e	p-value
1	30N	0	45	-0.93	0.23	0.5	0.5	0.0007
2	48VM	10	16	1.00	0.24	0.5	0.5	0.0007
3	50V	20	5	1.67	0.43	1.0	0.9	0.0009
4	54VA	11	84	0.62	0.16	0.8	0.6	0.0009
5	47V	10	17	1.03	0.30	0.5	0.5	0.0025
6	32I	10	21	0.85	0.26	0.5	0.5	0.0043
7	82AFST	20	100	0.46	0.16	0.7	0.6	0.0120
8	54LMST	11	36	0.54	0.19	0.3	0.3	0.0146
9	88S	0	9	-0.74	0.27	0.2	0.0	0.0179
10	73CSTA	2	66	0.52	0.22	0.6	0.6	0.0390
11	88DTG	0	44	-0.57	0.24	0.7	0.7	0.0390
12	33F	5	44	0.53	0.23	0.7	0.7	0.0419
13	24IF	2	16	0.67	0.31	0.1	0.1	0.0642
14	36ILVTA	0	141	0.30	0.15	1.0	0.5	0.0938
15	53LY	3	33	0.41	0.24	0.8	0.7	0.1410
16	84AV	11	73	0.22	0.17	0.5	0.5	0.3073
17	10FIRVY	2	217	-0.16	0.18	1.0	1.0	0.5457
18	20IMRTVL	0	115	0.14	0.15	1.0	0.9	0.5457
19	23I	0	4	0.68	1.02	0.3	0.0	0.6545
20	90M	10	171	0.13	0.19	1.0	1.0	0.6545
21	82MLC	10	4	0.39	0.62	0.2	0.2	0.6567
22	46ILV	11	143	0.06	0.16	1.0	0.8	0.8080
23	71TVI	2	181	0.06	0.15	1.0	1.0	0.8080
24	16E	0	9	-0.05	0.50	0.0	0.0	0.9308
25	63P	0	311	-0.03	0.17	0.6	0.6	0.9308
26	84C	10	2	0.15	1.74	0.0	0.0	0.9308

several of these mutations are adjusted for simultaneously that ETA problems arise.

Table 10: Other PI mutations not adjusted for among those mutations statistically significant at the 0.05 level. *Results are based on the non-parametric estimator using a targeted adjustment set selected based on the asymptotic criterion with* $p = 0.05$ *and* $B_{max} = 0.25$. *The effective adjustment set is selected data-adaptively. If the entry in a cell is empty, the variable importance estimate for the mutation in that row was adjusted for the mutation in that column. If the entry is not empty, the mutation in that column could not be adjusted for and the entry shows the sample correlation between the two relevant mutations.*

	30N	32I	47V	50V	54LMST	54VA	82AFST	84AV	88S
30N						-0.12	-0.20	-0.13	
32I			0.62		0.28		0.25		
47V		0.62			0.41		0.08		
48VM	-0.07			0.32	0.16	0.18	0.32		
50V									
54LMST		0.28	0.41			-0.14		0.21	
54VA							0.58		
82AFST	-0.20	0.25				0.58			
84AV	-0.13				0.21				
88S		-0.04	-0.03	-0.02		-0.08	-0.09	-0.03	

For the sake of illustration, figure 1 shows variable importance estimates for the two mutations 46ILV and 82AFST as a function of the adjustment level δ. For 46ILV, the algorithm selected a targeted adjustment level of $\delta^t = 1.0$. We note that the unadjusted estimate of 0.43 is significantly different from zero. As the adjustment set increases, however, the point estimates moves closer and closer to zero. At the same, we detect no appreciable ETA violation, suggesting the unadjusted effect is likely due to confounding. We also note that, as the adjustment increases, the standard errors for the corresponding point estimates first increase before finally dropping slightly below the level observed for the unadjusted estimate. The initial increase in efficiency relative to the unadjusted estimate is in agreement with recent findings by Moore and van der Laan (2007) according to which covariate adjustment in variable importance estimation typically results in increased precision if the ETA assumption is not violated.

For 82AFST, the algorithm selected a targeted adjustment level of $\delta^t = 0.7$. Again, the unadjusted estimate is significant, but increases in the adjustment set lead only to smaller corrections of the point estimate toward zero before a considerable ETA violation causes the estimates to become unstable for $\delta > 0.7$. We note that the efficiency relative to the unadjusted estimate again increases before the ETA violation causes it to drop sharply. These observations illustrate the dangers in performing a fully adjusted analysis in situations in which the ETA assumption is violated. We also note that the four different bias estimates displayed in the lower right panel are in good agreement with each other, again suggesting that the proposed algorithm is fairly robust with respect to the particular choices made for selecting the targeted adjustment set.

7 Discussion

In this paper, we propose a data-adaptive algorithm intended to increase the robustness of variable importance estimation with respect to violations of the ETA assumption. The algorithm is based on one of two identifiability criteria for selecting a targeted adjustment set as well as a mean-squared-error criterion for selecting an effective adjustment set. The data analysis shows very clearly the importance of selecting an appropriate targeted adjustment set as both unadjusted and fully adjusted analyses lead to unsatisfactory results. The fact that the algorithm chose not to adjust for some variables that have quite small marginal

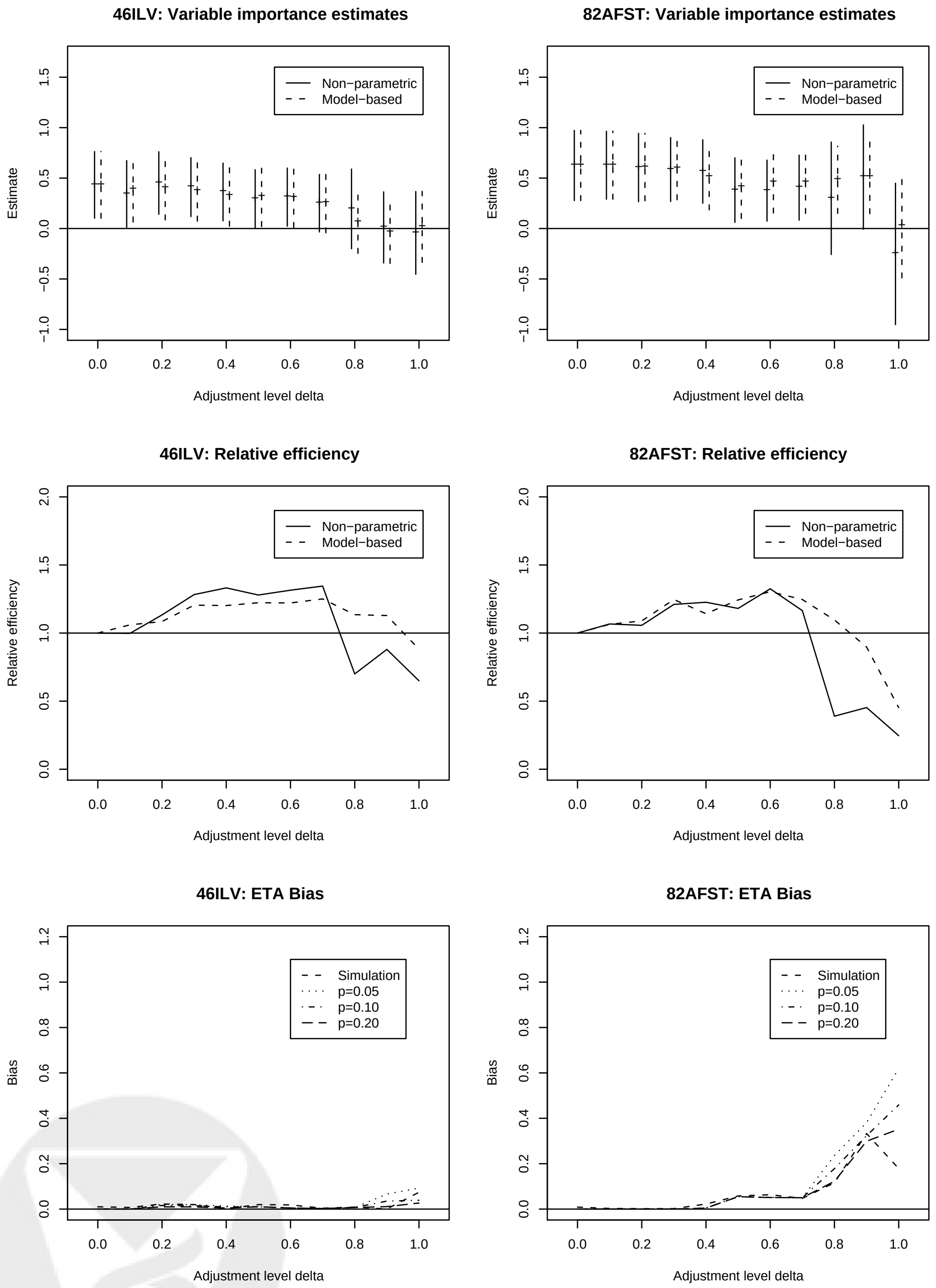

Figure 1: Variable importance estimates for the two mutations 46ILV and 82AFST as a function of the adjustment level δ. *In the first row, variable importance estimates along with 95% confidence intervals are shown. In the second row, relative efficiencies as compared to the unadjusted variable importance estimate are shown. In the third row, a simulation-based finite-sample ETA bias estimate is shown along with asymptotic bias estimates for IPTW estimators truncated such that no weight is greater than a proportion p of the sum of all weights.*

correlations with the mutation of interest suggests that serious practical ETA violations may be much more common than previously thought and underscore the need to assess the validity of this assumption. This point is particularly important since many conventional approaches to biomarker discovery such as regression analysis typically do not reveal such problems through sharply inflated standard errors as seen with the non-truncated IPTW and targeted maximum-likelihood estimator, thus not giving the investigator any warning that the parameter of interest may be poorly identified from the observed data.

The data analysis and the simulation study also illustrate the potential gains in efficiency that can be achieved by selecting the effective adjustment set data-adaptively. In the data analysis, the proposed algorithm for selecting both adjustment sets data-adaptively identified a subset of mutations that is in excellent agreement with current understanding of lopinavir resistance, in better agreement, in particular, than previous analyses that categorically excluded other candidate PI mutations from the adjustment set. These findings suggest that variable importance estimation based on data-adaptive selection of the targeted and effective adjustment sets represents a promising new approach for studying the effects of HIV mutations on clinical virologic response to antiretroviral therapy as well as for biomarker discovery in general.

A number of possible extensions of the methodology discussed in this article exist. First, the approach can be applied in a straightforward way to the estimation of causal effects in the point-treatment setting by use of marginal structural models (Robins et al., 2000). In addition, the methodology can be extended to longitudinal data structures. In that case, selection of the targeted adjustment set would need to be based on the parametric-bootstrap approach for estimating the finite-sample bias of the IPTW estimator since closed-form asymptotic bias estimates for truncated IPTW estimators are not currently available for the longitudinal setting. Selection of the effective adjustment set would still be based on G-computation point estimates that now, however, would also need to be obtained through a Monte-Carlo simulation. Some care would need to be taken in defining a nested sequence of candidate adjustment sets for each time point. It might be preferable to not consider different candidate adjustment sets for different time points, but instead to define identical candidate adjustment sets across time points.

In face of the small marginal correlations between some of the treatment variables and other candidate PI mutations that could not be adjusted for, it might be useful to explore alternative criteria for defining the sequence of candidate adjustment sets. This may be less important in cases in which the ETA assumption is satisfied, but once the adjustment set is large enough to result in an appreciable violation, it might be advantageous to add covariates to the adjustment set directly based on the effect that they would have on the identifiability criterion.

As mentioned in section 4, inference based on the influence curve may be somewhat optimistic in finite samples if the effective adjustment set is selected data-adaptively. Future research is needed to compare inference based on this approach to inference based on an honest bootstrap procedure to quantify the extent to which the use of the former might be problematic. We note, however, that even in situations in which such p-values may be systematically optimistic, they would be still be useful for obtaining a meaningful ranking of the candidate biomarkers.

References

Bembom, O., Petersen, M., Rhee, S.-Y., Fessel, W., Sinisi, S., Shafer, R., and van der Laan M.J. (2007). Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant hiv infection. Technical Report 221, UC Berkeley Division of Biostatistics Working Paper Series.

Bembom, O. and van der Laan, M. (2007a). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report XXX, UC Berkeley Division of Biostatistics Working Paper Series.

Bembom, O. and van der Laan, M. (2007b). Estimating the effect of vigorous physical activity on mortality in the elderly based on realistic individualized treatment and intention-to-treat rules. Technical Report 217, UC Berkeley Division of Biostatistics Working Paper Series.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statstical Society: Series B*, 57:289–300.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Moore, K. and van der Laan, M. (2007). Covariate Adjustment in Randomized Trials with Binary Outcomes: Targeted Maximum Likelihood Estimation. Technical Report 215, UC Berkeley Division of Biostatistics Working Paper Series.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference*, 129:405–426.

Rhee, S., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103:17355–17360.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Shafer, R., Jung, D., and Betts, B. (2000). Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nature Medicine*, 6(11):1290–1292.

Sinisi, S. and van der Laan, M. (2004). Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Appliations in Genetics and Molecular Biology*, 3(1):Article 18.

van der Laan, M. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):Article 2.

van der Laan, M. and Robins, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* Springer Series in Statistics. Springer Verlag.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

Wang, Y., Petersen, M., Bangsberg, D., and van der Laan, M. (2006). Diagnosing Bias in the Inverse-Probability-of-Treatment-Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. Technical Report 211, UC Berkeley Division of Biostatistics Working Paper Series.