STATISTICS IN MEDICINE
Statist. Med. (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3414

Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant HIV infection

Oliver Bembom[1,*,†], Maya L. Petersen[1], Soo-Yon Rhee[2], W. Jeffrey Fessel[3], Sandra E. Sinisi[1], Robert W. Shafer[2] and Mark J. van der Laan[1]

[1]*Division of Biostatistics, University of California, Berkeley, CA, U.S.A.*
[2]*Division of Infectious Diseases, Center for AIDS Research, Stanford University, Palo Alto, CA, U.S.A.*
[3]*Clinical Trials Unit, Kaiser Permanente, San Francisco, CA, U.S.A.*

SUMMARY

Researchers in clinical science and bioinformatics frequently aim to learn which of a set of candidate biomarkers is important in determining a given outcome, and to rank the contributions of the candidates accordingly. This article introduces a new approach to research questions of this type, based on targeted maximum-likelihood estimation of variable importance measures.

The methodology is illustrated using an example drawn from the treatment of HIV infection. Specifically, given a list of candidate mutations in the protease enzyme of HIV, we aim to discover mutations that reduce clinical virologic response to antiretroviral regimens containing the protease inhibitor lopinavir. In the context of this data example, the article reviews the motivation for covariate adjustment in the biomarker discovery process. A standard maximum-likelihood approach to this adjustment is compared with the targeted approach introduced here. Implementation of targeted maximum-likelihood estimation in the context of biomarker discovery is discussed, and the advantages of this approach are highlighted. Results of applying targeted maximum-likelihood estimation to identify lopinavir resistance mutations are presented and compared with results based on unadjusted mutation–outcome associations as well as results of a standard maximum-likelihood approach to adjustment.

The subset of mutations identified by targeted maximum likelihood as significant contributors to lopinavir resistance is found to be in better agreement with the current understanding of HIV antiretroviral resistance than the corresponding subsets identified by the other two approaches. This finding suggests that targeted estimation of variable importance represents a promising approach to biomarker discovery.

KEY WORDS: biomarker discovery; variable importance; targeted maximum-likelihood estimation; HIV drug resistance

*Correspondence to: Oliver Bembom, Division of Biostatistics, University of California, Berkeley, CA, U.S.A.
†E-mail: bembom@berkeley.edu

Received 2 August 2007
Accepted 21 July 2008

1. INTRODUCTION

Researchers in bioinformatics, biostatistics, and related fields are often faced with a large number of candidate biomarkers and aim to assess their importance in relation to a given outcome. Examples include the identification of single nucleotide polymorphisms associated with the development of cancers, identification of HLA types associated with disease progression rates, and the identification of viral mutations that contribute to reduced susceptibility to drug therapy. In some cases, the goal may be to select from a list of candidates those biomarkers with underlying mechanistic relationships to the outcome. In others, the researcher may wish to rank the importance of a set of candidate biomarkers in terms of their contributions to determining the outcome.

In this article we introduce a novel method for biomarker discovery based on targeted maximum-likelihood estimation of variable importance measures (VIMs) [1]. As we discuss, the marginal association of a candidate biomarker with the outcome may not reflect the biomarker's mechanistic or prognostic significance. For example, a viral mutation may be associated with poor response to a given drug without playing any mechanistic role in resistance, as a result of covariates that both predict the presence of the mutation and affect the outcome via an alternative pathway. VIMs provide a means to rank candidate biomarkers based on their association with a given outcome, controlling for a large number of additional covariates [2]. Specifically, given a binary candidate biomarker A, an outcome Y, and a list of covariates W, the W-adjusted VIM is defined as $E_W(E(Y|A=1,W)-E(Y|A=0,W))$.

Several approaches are available to estimate VIMs. Perhaps the most common approach is based on maximum-likelihood estimation of the conditional expectation of the outcome given the candidate biomarker and covariates. This conditional expectation is then evaluated at $A=1$ and 0 for each subject, and the difference is averaged across the population. Such an approach corresponds to the G-computation formula of Robins [3] applied at a single time point.

In this article, we show how a recent advance in statistical methodology, targeted maximum-likelihood estimation, can improve on this standard approach. Targeted maximum-likelihood estimation involves a simple one-step adjustment to an initial estimate of the conditional expectation of the outcome given the biomarker and covariates. This adjustment reduces the bias in the estimate of the VIM and improves the robustness to mis-specification of the likelihood. The theoretical basis for targeted maximum-likelihood estimation was recently published by van der Laan and Rubin [1]. Here, we demonstrate how this work can be applied in practice to improve standard approaches to biomarker discovery. Throughout the article, emphasis is placed on practical understanding and implementation of the methods described.

Targeted maximum likelihood is illustrated using an original data example drawn from the treatment of antiretroviral-resistant HIV infection. Using observational clinical data, we aimed to determine which of a set of candidate viral mutations affects clinical virologic response to the antiretroviral drug lopinavir, and to rank the importance of these mutations for drug-specific resistance. The resulting ranking can be used to inform interpretation of viral genotypes, and to aid clinicians in selecting new antiretroviral treatment regimens with a greater probability of virologic success.

1.1. Outline

The article has the following structure. Section 2 introduces the data application and provides background on the research question and the data structure. In Section 3, we discuss methods

DOI: 10.1002/sim

for biomarker discovery, and compare estimation of unadjusted and adjusted associations between the candidate biomarker and the outcome ($E(Y|A=1)-E(Y|A=0)$ and $E_W(E(Y|A=1,W)-E(Y|A=0,W))$, respectively). Section 4 presents the targeted maximum-likelihood approach to the estimation of W-adjusted VIMs, and compares it with a standard (or G-computation) approach. Implementation and inference using the targeted approach are discussed both generally and in the context of the data example. Section 5 presents the results of the data analysis, in which the importance of candidate mutations was assessed using unadjusted, G-computation, and targeted estimates of VIMs. We compare the results of these methods and discuss them in the context of the current understanding of HIV antiretroviral resistance. Section 6 concludes with a discussion.

2. APPLICATION: IDENTIFICATION OF HIV MUTATIONS ASSOCIATED WITH DECREASED VIRAL SUSCEPTIBILITY TO LOPINAVIR

2.1. Research question

Virus resistant to antiretroviral drugs frequently evolves during treatment of HIV infection and can result in disease progression if new therapies are not initiated. Designing an effective salvage therapy regimen for an individual infected with resistant virus requires choosing drugs to which the virus infecting that individual remains sensitive. Tests of viral resistance are now available to help guide salvage regimen design. However, interpretation of the results of resistance tests for the purposes of guiding salvage regimen drug choice remains complex.

Assays of viral susceptibility to antiretroviral drugs fall into two general categories: phenotype-based and genotype-based. Phenotypic resistance tests directly quantify *in vitro* drug susceptibility using recombinant virus, while genotypic resistance tests are performed by sequencing the genes for the viral protease and reverse transcriptase enzymes, the targets of the major antiretroviral classes. While genotypic tests are less expensive, less complex, and faster to perform than phenotypic tests, interpretation of the results of genotypic tests requires linking patterns of viral mutations to *in vivo* and *in vitro* resistance.

Data from several sources have been used to inform interpretation of viral genotype. Observed associations between the presence of specific viral mutations and patients' treatment histories suggest that these mutations have been selected for over the course of therapy and likely contribute resistance to the specific drugs used. *In vitro* experiments have also provided insight into the role of individual mutations in determining drug-specific viral susceptibility. Such experiments include observation of viral evolution in the presence of antiretroviral drugs and tests of the ability of mutated viruses to replicate in the presence of drug. The resulting data on links between viral mutations and susceptibility to antiretroviral drugs have been combined to create rule-based algorithms for the interpretation of genotype data. Examples include the French ANRS (National Agency for AIDS Research) algorithm [4], the Rega algorithm [5], and the Stanford HIVdb program [6]. The Stanford algorithm in particular provides drug-specific estimates of viral susceptibility using a weighted scoring system for mutations thought to be associated with resistance. Viral susceptibility to an entire regimen is calculated by summing susceptibility scores for each drug in the regimen, yielding a genotypic susceptibility score (GSS). The International AIDS Society also publishes an annual drug-specific list of mutations thought to affect viral resistance [7].

Ultimately, the goal of such algorithms is to identify mutations with large impacts on clinical drug response. We aimed to use data from an observational clinical cohort to rank a list of candidate

DOI: 10.1002/sim

resistance mutations based on their importance in conferring resistance to specific antiretroviral drugs. For the sake of illustration, we focused on resistance to the commonly used protease inhibitor (PI) drug lopinavir. Rankings like the one presented here can be used to inform current genotype interpretation algorithms, with the aim of improving selection of salvage antiretroviral drug regimens for patients infected with resistant HIV virus.

2.2. Data

2.2.1. Study sample and inclusion criteria. Analyses were based on 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 6000 patients. Data collected include use of antiretroviral drugs, results of viral genotype tests, and measurements of plasma HIV RNA level (viral load) and CD4 T cell count collected during the course of clinical care.

We identified all treatment change episodes (TCEs) in this database, which 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 the 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.

2.2.2. Data structure. Baseline genotype was summarized as a vector $\mathbf{A}$ of binary variables A_j that indicate the presence of a specific mutation in the protease enzyme of HIV (the viral target of lopinavir). 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 on 18 July 2006). In total, we considered 30 candidate PI mutations. In the sections that follow, we describe methods for estimating the importance of a single candidate biomarker A. In applying these methods to the data example, each of the candidate mutation A_j, for $j=1,\ldots,30$, was assessed separately; however, for simplicity we suppress the subscript j.

Antiretroviral regimens generally combine drugs from more than one class. The following characteristics of the non-PI component of the salvage regimen were included in the set W of adjustment variables: indicators of use of each of 13 non-PI drugs; number of drugs used in each major non-PI class (nucleoside reverse transcriptase inhibitors or NRTI and non-nucleoside reverse transcriptase inhibitors or NNRTI); 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 regimens.

DOI: 10.1002/sim

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.

Summaries of non-PI mutations in the baseline genotype (i.e. mutations in the reverse transcriptase enzyme targeted by the NRTI and NNRTI classes) were also included in the covariate set W. Known NRTI and NNRTI resistance mutations present at baseline were summed. In addition, 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 GSS, which summarized the activity of the non-PI component of the regimen.

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 the final $\log_{10}$ viral load over the 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 the $\log_{10}$ 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 the viral load detected in the population, which was -4.2 log. Under this definition, both large drops in the 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.

In summary, each TCE contained a baseline viral genotype, summarized in a vector $\mathbf{A}$ of binary variables defining the presence or absence of each of a list of candidate PI resistance mutations, a new antiretroviral regimen containing lopinavir initiated following the genotype, and an outcome Y capturing the change in the $\log_{10}$ viral load at 4–36 weeks (measured before any subsequent changes in regimen) over the baseline $\log_{10}$ viral load. In addition, each TCE contained a set W of adjustment variables, which included summaries of the non-PI mutations in the viral genotype, as well as covariates collected both prior to and following the genotype. We aimed to rank the candidate PI mutations based on their impact on clinical outcome. In the sections that follow, we discuss several general approaches to research questions of this type and discuss their implementation in the context of this data example.

3. BACKGROUND: STATISTICAL METHODS FOR BIOMARKER DISCOVERY

3.1. Marginal versus adjusted biomarker–outcome associations

One straightforward approach to biomarker discovery is to assess the unadjusted association between each candidate biomarker and the outcome or, in other words, to estimate $E(Y|A=1) - E(Y|A=0)$ for each candidate A. In some settings the unadjusted association may be the quantity of interest, particularly when biomarkers can be experimentally manipulated. For example,

DOI: 10.1002/sim

if the researcher is able to induce specific mutations in a virus without altering other key covariates and then to compare viral replication in the presence and absence of each mutation, then the assessment of marginal associations may be an appropriate approach.

In others settings, however, the marginal association between a candidate biomarker and the outcome can be misleading or fail to capture the underlying mechanistic relationship of interest. When dealing with observational or clinical data, covariates are often present that are both associated with the candidate biomarker and also affect the outcome via a pathway independent of the biomarker. Such covariates are known in the epidemiologic literature as confounders.

The HIV data example illustrates how confounding of a biomarker effect can occur. HIV-infected patients with a given mutation may disproportionately include subjects with an extensive treatment history. Because past treatment can strongly affect the presence of other mutations, past treatment patterns can cause a viral mutation with no effect on resistance to occur commonly with mutations that do strongly affect resistance. The candidate mutation may thus appear to confer resistance when in fact it is simply acting as a marker for past treatment history and the presence of other mutations. The picture is further complicated by the fact that in HIV infection, past mutations can be 'archived' and remain present only in latent virus. Such archived mutations are not observable, but can still impact clinical response. We aimed to capture information about these archived mutations via covariates describing a subject's treatment history prior to initiation of the salvage regimen. In the HIV application, then, controlling for the presence of other mutations and for past treatment history would allow us to isolate to what extent any decreased virologic response we observe is due to the presence of the candidate mutation being considered.

The W-adjusted VIM $E_W(E(Y|A=1,W)-E(Y|A=0,W))$ removes the confounding effect of other covariates W by comparing the effect of A on Y in subgroups of patients with identical values w of these covariates, i.e. by looking at $E(Y|A=1,W=w)-E(Y|A=0,W=w)$. The VIM is then obtained by simply averaging such comparisons over the marginal distribution of W. Adjustment for W requires, however, that A shows sufficient variation in all strata of W. The VIM of a given mutation could not be adjusted for treatment history, for example, if the mutation *always* occurred among subjects with a specific treatment history; in this case there is clearly not enough information in the data to estimate the difference in clinical response that would be seen in the presence versus absence of the mutation in this sub-population. In the data example, the candidate PI mutations were highly collinear; as a result, for a given candidate mutation, we were unable to adjust for the presence of the other candidate PI mutations.

3.2. Adjustment for post-biomarker covariates

Selecting which covariates to adjust for when estimating the VIM requires careful thought and substantial background knowledge about the specific data application to which the method is being applied. We discussed above the need in the HIV data example to control for at least two types of baseline covariates, treatment history prior to salvage regimen initiation and the presence of non-PI mutations. However, in some settings it may also be desirable to adjust for covariates that occur after, and may be affected by, the candidate biomarker of interest.

In the HIV data example, the non-PI drugs contained in the salvage regimen, assigned after assessment of viral genotype, may differ according to the presence of a candidate mutation. Such a scenario could arise, for example, if the clinician observed a mutation known to result in high-level resistance, and in response increased the potency of the subject's background (non-PI) regimen. To the extent that differences in background regimen impact clinical response, they have the potential

DOI: 10.1002/sim

to obscure drug resistance caused by the candidate mutation. In the causal inference framework, this scenario can be viewed as a (spurious) indirect effect of the mutation. Our aim is to estimate the direct effect of the mutation on clinical response, blocking any possible effect the presence of the mutation might have on the clinician's choice of background salvage regimen.

One option is to simply include post-biomarker covariates together with baseline covariates in the covariate set W. However, interpretation of the resulting W-adjusted VIM requires careful thought in the context of the specific data example to which it is being applied. Let W_b denote baseline covariates (occurring prior to the biomarker A), and let Z denote covariates occurring after, and affected by, A. At an individual level, the quantity $E(Y|A=1, Z=z, W_b) - E(Y|A=0, Z=z, W_b)$ corresponds (under assumptions on confounders—see [8]) to the effect of the biomarker on the outcome holding the intermediate variables Z at a fixed level. The mean of these individual effects provides a population summary: $E_{W_b}(E(Y|A=1, Z=z, W_b) - E(Y|A=0, Z=z, W_b))$. In the HIV example, this quantity would correspond with estimating the mean difference in virologic response if the researcher induced a candidate mutation to be present versus absent, and assigned a salvage regimen with fixed characteristics regardless of the presence of the mutation.

If one is willing to assume the absence of interaction between A and Z, then

$$\begin{aligned} &E_{W_b}(E(Y|A=1, Z=z, W_b) - E(Y|A=0, Z=z, W_b)) \\ &\quad = E_{ZW_b}(E(Y|A=1, W_b, Z) - E(Y|A=0, W_b, Z)) \end{aligned} \tag{1}$$

In other words, averaging over the empirical distribution of the post-biomarker covariates, Z, will not alter the estimated VIM, and thus the direct effect of interest can be estimated by simply including post-biomarker covariates together with baseline covariates in the adjustment set W. In the HIV example, the no-interaction assumption corresponds with assuming that the effect (or adjusted VIM) for each candidate PI mutation does not differ depending on the characteristics of the background regimen, a reasonable assumption given that PI mutations are not expected to affect response to non-PI drugs. In the analyses reported, characteristics of the (non-PI) background regimen were therefore included in the adjustment set W.

An additional common post-biomarker covariate is the duration between the assessment of the biomarker and the measurement of the outcome. To the extent that this duration is variable, differs depending on the presence of the biomarker, and affects the outcome, it has the potential to obscure the VIM of interest. In the HIV example, the outcome viral load was assessed between 4 and 36 weeks following salvage regimen initiation, and viral loads observed sooner following salvage initiation were likely to be higher. If the presence of a candidate mutation affected the time at which the viral load was monitored, duration until the outcome was monitored could thus serve as an additional source of a spurious indirect effect. In the analyses reported in this article, time until the viral load assessment was included as a covariate in W, according to the following rationale: (1) If the presence of the candidate mutation did not affect duration until the outcome assessment, this duration could not serve as a source of an indirect effect, and the inclusion of duration as a covariate did not require any additional assumptions; however, given the association between the duration and the outcome, the inclusion of this covariate would be expected to improve the efficiency. (2) If the presence of the candidate mutation did affect duration until the outcome assessment, we wished to control for this indirect effect; inclusion of duration as a covariate allowed us to do this, again under the no-interaction assumption (interpretable in this case as assuming that the effect of the mutation on virologic response did not vary over time). We note that inclusion of duration until the outcome assessment is one possible way to address a potentially informative

DOI: 10.1002/sim

censoring mechanism; alternatives, such as the use of inverse probability weights [9], are beyond the scope of this article.

In summary, depending on the data application, inclusion of post-biomarker covariates in the adjustment set W may be warranted. However, such a decision requires a careful consideration of the interpretation of the resulting W-adjusted VIM. In the following section, we return to the estimation of this parameter.

3.3. A traditional approach to the estimation of VIMs

A common approach to the estimation of W-adjusted VIMs focuses on estimation of the conditional expectation $E(Y|A, W)$ of the outcome given the biomarker and covariates, using standard maximum-likelihood estimation. Given an estimate of $E(Y|A, W)$, the VIM can be estimated by simply evaluating this object at the values $A = 0$ and 1, and averaging the resulting differences across the population. Such an approach of intervening on the likelihood corresponds to the G-computation formula of Robins [3], applied in the setting of a single time point. Frequently, the number of covariates W is large and the functional form of $E(Y|A, W)$ is unknown. Multiple algorithms are available to learn this form data-adaptively; examples include classification and regression trees [10], random forests [11], least angle regression [12], and the Deletion/Substitution/Addition (D/S/A) algorithm [13]. Either cross-validation or some form of penalization of the likelihood is generally used to select the level of model complexity providing the optimal bias–variance trade-off for the purposes of prediction; in the case that Y is continuous, this corresponds to selecting the level of complexity, that minimizes the mean-squared error.

Such an approach is appropriate if the goal of the analysis is to find the optimal predictor of the outcome Y given A and W. However, biomarker discovery often aims instead to evaluate a list of candidate biomarkers, rank them in terms of importance, and identify those significantly associated with the outcome. When the goal of the analysis is to estimate the W-adjusted VIM for each of the candidate biomarkers, a different estimation approach may be warranted. To understand why, consider the HIV data example.

The number of covariates in this application, as in many biomarker applications, is very large, consisting of multiple mutations, salvage regimen characteristics, baseline characteristics of the subject such as viral load and CD4 count, and the subject's past antiretroviral treatment experience. A conventional approach would attempt to choose the model that best predicts virologic response as a function of the candidate mutation and these covariates. Given the large number of covariates, a reasonable approach would be to apply some data-adaptive regression algorithm to select this model. However, standard data-adaptive approaches aim to achieve the optimal bias–variance trade-off for the entire conditional expectation of Y given A and W. Because the VIM is a much smoother parameter, a model fit for the purpose of prediction will generally not provide the best bias–variance trade-off for the purpose of estimating the VIM. Furthermore, a predictor constructed using conventional methods is likely to involve multiple terms that do not contain the candidate mutation; for example, the baseline viral load and the CD4 T cell count are likely to make important contributions to virologic response regardless of the mutation profile. Mis-specification of such terms in, for example, a traditional multivariable regression model can result in bias in the estimated effect of the mutation, even under the null hypothesis of no mutation effect.

In summary, in the context of biomarker discovery, prediction is often not the underlying goal of analysis. Traditional approaches invest in achieving a good fit for the entire conditional expectation of Y given A and W; however, such a fit is not targeted at the biomarker-specific VIM of interest. In

DOI: 10.1002/sim

contrast, *targeted* maximum-likelihood estimation of the VIM, introduced in the following section, allows the researcher to focus on the importance of each mutation, in turn reducing the bias in the adjusted VIM estimate and improving the robustness to mis-specification of the model for $E(Y|A,W)$.

4. TARGETED MAXIMUM-LIKELIHOOD ESTIMATION

In this section, we provide a practical overview of targeted maximum-likelihood estimation of VIMs. The formal statistical theory behind targeted maximum likelihood has been published elsewhere [1]. Here, our aim is to make this material practically accessible to the practitioner who wishes to apply targeted maximum-likelihood estimation to improve biomarker discovery.

The density of the observed data $O=(W,A,Y)$ is defined by the marginal distribution of covariates W, the conditional distribution $P(A|W)$ of the biomarker given covariates, and the conditional distribution $P(Y|A,W)$ of the outcome Y given A and W. Unlike standard approaches to VIM estimation (which rely entirely on estimating $E(Y|A,W)$), targeted maximum-likelihood estimation also involves estimation of $P(A|W)$. This estimate of the conditional distribution of the biomarker given covariates is used to update an initial estimate of $E(Y|A,W)$ in a way that targets that estimate at the goal of estimating the W-adjusted VIM of the treatment variable A of interest. This targeting step, followed by the evaluation of the updated estimate at $A=1$ and 0 and taking the empirical mean, results in an estimator with reduced bias and improved robustness to model mis-specification.

In the context of classic parametric models, a single maximum-likelihood estimate of the entire data-generating distribution is sufficient to answer equally well any question about different parameters of that distribution. It holds true, for example, that the maximum-likelihood estimate of any function f of a particular parameter θ is equal to that function f of the maximum-likelihood estimate of θ. For the sake of maximum-likelihood estimation it therefore does not matter if the goal is to estimate θ of $f(\theta)$. As soon as one avoids relying on the often unrealistic assumptions corresponding to parametric models, however, it is no longer possible to obtain a single estimate of the entire data-generating distribution, which can then be used to extract estimates of any parameter of that distribution one might be interested in. Instead it becomes necessary to target the estimation procedure at the particular parameter of interest in order to obtain reliable results [14]. It is this avoidance of parametric modeling assumptions throughout this article that motivates the targeting step that lies at the hard of targeted maximum-likelihood estimation.

We now denote our parameter of interest, the W-adjusted VIM, by

$$\theta \equiv E_W[E(Y|A=1,W)-E(Y|A=0,W)] \tag{2}$$

To ensure that this parameter is well-defined, we will assume that

$$0<P(A=1|W)<1 \tag{3}$$

with probability one or, in other words, that some variation in the biomarker exists within each stratum of W. As mentioned previously, insufficient variation of A in the strata of W causes the parameter of interest to become hard to identify from the observed data. We note that this problem is not a feature of targeted maximum-likelihood estimation, but rather of the parameter of interest so that conventional approaches for estimating the W-adjusted VIM rely on assumption (3)

DOI: 10.1002/sim

as well. In traditional regression techniques, in fact, it has long been recognized that individual coefficient estimates can become highly variable if several explanatory variables are strongly correlated, a problem that is typically referred to as collinearity. In our data example, we addressed the high correlation between the different candidate PI mutations by not adjusting the VIM of a given candidate PI mutation for the presence of other candidate PI mutations. More sophisticated approaches for dealing with such issues have recently been proposed [15, 16], but are beyond the scope of this article.

We first summarize the basic steps involved in targeted maximum-likelihood estimation of θ before going on to discuss each in detail, illustrated in the context of the data example. Implementation of the targeted maximum likelihood involves the following steps:

1. Estimate the conditional expectation of Y given A and W. We denote this initial estimate as $Q_n^0(A, W)$.
2. Estimate the conditional distribution of the biomarker given covariates. We denote this estimate as $g_n^0(A, W)$.
3. For each subject, calculate a specific covariate, based on the subject's observed values for A and W and using the estimate $g_n^0(A, W)$. We denote this covariate as $h(A, W)$.
4. Update the initial regression $Q_n^0(A, W)$ by adding the covariate $h(A, W)$ and estimating the corresponding coefficient by maximum likelihood, holding the remaining coefficient estimates fixed at their initial values. We denote this updated regression as $Q_n^1(A, W)$.
5. Evaluate the updated regression at $A = 1$ and 0 to get two predicted outcomes for each subject and taking the empirical mean of the difference across the population to obtain a targeted estimate of the VIM.

4.1. An initial estimate of $E(Y|A, W)$

The first step in targeted maximum-likelihood estimation consists of obtaining an initial estimate of the conditional expectation $E(Y|A, W)$ of Y given A and W, as one would do in a standard G-computation approach to variable importance estimation. The number of covariates W will often be large, and the functional form for $E(Y|A, W)$ will often be unknown. In this case, as discussed in Section 3.3, a range of data-adaptive approaches are available to obtain an estimate $Q_n^0(A, W)$.

In the HIV data example, we were faced with a large number of candidate covariates, detailed in Section 2.2. These included mutations other than the candidate mutation of interest (incorporated both as individual covariates and summarized using measures such as drug-specific susceptibility scores), various summaries of the past treatment history, baseline laboratory data on CD4 T cell count and viral load, time until outcome assessment, and summary measures of the background regimen and its estimated activity given baseline genotype. To reduce the size of the adjustment set W, we first performed a dimension reduction based on the unadjusted association of each candidate covariate with the outcome Y; the covariates with the 50 smallest p-values were retained.

Following this dimension reduction, we applied the D/S/A algorithm [13] to obtain an initial estimate $Q_n^0(A, W)$ based on the remaining 50 covariates. The D/S/A algorithm is a data-adaptive algorithm for polynomial regression, which generates candidate predictors as linear combinations of polynomial tensor products in continuous and/or binary covariates. These candidate estimators are indexed by the number and complexity of the terms, and the optimal candidate is selected using cross-validation. In estimating $E(Y|A, W)$, the D/S/A algorithm considered candidate estimators with up to two-way interaction terms and a maximum quadratic order for each term. Specifically, $E(Y|A, W)$ was modeled by first selecting a model for $E(Y|W)$ with a maximum of 10 terms, then

DOI: 10.1002/sim

adding the term A to the selected model, and finally re-running the algorithm to select a model for $E(Y|A,W)$, forcing previous terms to be in the model and allowing the D/S/A algorithm to add up to 5 new terms. In estimating the VIM of mutation p10FIRVY, for example, the D/S/A algorithm selected the linear regression model

$$E[Y|A,W] = \beta_0 + \beta_1 \text{p10FIRVY} + \beta_2 \text{IDV} + \beta_3 \text{LPV} + \beta_4 \text{BLVL} + \beta_5 \text{GSS} + \beta_6 \text{r184V} + \beta_7 \text{r106M} + \beta_8 \text{BLVL}^2 + \beta_9 \text{r67STH} + \beta_{10} \text{r103NST} + \beta_{11} \text{r98G}$$

where r67STH, r103NST, r106M, and r184V are mutations in the reverse transcriptase gene of HIV; IDV and LPV are indicators for having previously been treated with indinavir and lopinavir; BLVL is baseline viral load; and GSS is the genotypic susceptibility score. This initial estimate of $E(Y|A,W)$ was evaluated at $A=1$ and 0, and the empirical mean of the difference was used to estimate VIMs according to the G-computation approach. In other words, the G-computation estimate of the VIM was given by

$$\theta_n^{\text{G-comp}} = \frac{1}{n} \sum_{i=1}^{n} Q_n^0(1, W_i) - Q_n^0(0, W_i) \tag{4}$$

The targeted maximum-likelihood estimate of the VIM also made use of this initial estimate Q_n^0, updated according to the following steps.

4.2. Estimation of $P(A|W)$

The next step in the targeted estimation of VIMs consists of estimating the conditional distribution of A given W. In the current application, A is binary so that a logistic regression model can be used for this purpose. In fitting such a model, we first employed the same dimension reduction on W as used in fitting $E(Y|A,W)$. We then used the D/S/A algorithm to data-adaptively select an appropriate logistic regression model for the probability of having the candidate mutation given W. The D/S/A algorithm was run with a maximum of two-way interactions, a maximum quadratic order for each term, and a maximum of 10 terms. In estimating the VIM of mutation p10FIRVY, for example, the D/S/A algorithm selected the logistic regression model

$$\log\left(\frac{P(A|W)}{1-P(A|W)}\right) = \beta_0 + \beta_1 \text{r41L} + \beta_2 \text{APV} + \beta_3 \text{r67EGN} + \beta_4 \text{r184V} + \beta_5 \text{SQV}$$

where r41L, r67EGN, and r184V are mutations in the reverse transcriptase gene of HIV and APV and SQV are indicators for having previously been treated with amprenavir and saquinavir. The practical performance of the targeted maximum-likelihood estimator can be improved somewhat by ensuring that no estimated treatment probabilities $g_n^0(A,W)$ are very close to zero; here, we did so by setting estimated treatment probabilities smaller than 0.01 to 0.01.

4.3. Calculation of $h(A,W)$ and update of $Q_n^0(A,W)$

Using the resulting estimate $g_n^0(A,W)$, the next step is to calculate the following covariate, denoted as $h(A,W)$, for each subject:

$$h(A,W) \equiv \left(\frac{I(A=1)}{g_n^0(1,W)} - \frac{I(A=0)}{g_n^0(0,W)}\right) \tag{5}$$

DOI: 10.1002/sim

A one-step adjustment to the initial regression estimate $Q_n^0(A, W)$ is performed by adding the covariate $h(A, W)$ to this regression and obtaining a maximum-likelihood estimate ε_n of the corresponding coefficient ε, holding all other coefficient estimates fixed at their initial values. The estimate ε_n can thus be obtained by regressing Y on $h(A, W)$ using $Q_n^0(A, W)$ as an offset. The updated estimate $Q_n^1(A, W)$ is then given by

$$Q_n^1(A, W) = Q_n^0(A, W) + \varepsilon_n h(A, W) \tag{6}$$

The corresponding targeted estimate of the marginal VIM is given by

$$\theta_n^{\text{T-MLE}} = \frac{1}{n} \sum_{i=1}^{n} Q_n^1(1, W_i) - Q_n^1(0, W_i) \tag{7}$$

The targeted maximum-likelihood estimator is thus identical to the G-computation estimator described above except that it is based on the updated regression fit $Q_n^1(A, W)$ rather than the initial fit $Q_n^0(A, W)$.

Standard errors estimates and p-values for the targeted maximum-likelihood VIM estimator can be obtained using the non-parametric bootstrap. This approach provides a straightforward means to address the dependence between observations, as occurred in the data example, because a single subject could contribute more than one TCE to the analyses. The non-parametric bootstrap also offers an opportunity to perform re-sampling-based approaches to multiple testing without substantial additional computer time.

4.4. Advantages of targeted maximum-likelihood estimation

Standard approaches to the estimation of variable importance rely entirely on the estimation of the conditional expectation of the outcome given the biomarker and covariates. The approach presented here provides a means to target this regression estimate specifically at the parameter of interest (in this case the W-adjusted VIM). In the context of the HIV data, for example, targeted maximum-likelihood estimation of W-adjusted variable importance allows us to obtain a targeted estimate of the significance of each candidate resistance mutation in turn.

If the initial estimate of $E(Y|A, W)$ is based on standard multivariable or logistic regression, implementing the targeted maximum-likelihood estimator is simply a matter of adding a covariate to the initial regression and estimating the corresponding coefficient by maximum likelihood. The result of this single-step adjustment is an improved robustness to model mis-specification in comparison with a G-computation estimate based on the initial regression fit [1]. Specifically, the G-computation estimator is consistent only if $E[Y|A, W]$ is estimated consistently. If a model for $E[Y|A, W]$ is specified *a priori*, then the G-computation estimator is consistent only if that model is correct. If the true dependence of Y on A and W is given by

$$E[Y|A, W] = -2 + 10AW_1W_2 - 8AW_3W_4 \tag{8}$$

for example, but the investigator relies on the model

$$E[Y|A, W] = \beta_0 + \beta_1 A + \beta_2 W_1 + \beta_3 W_2 + \beta_4 W_3 + \beta_5 W_4 \tag{9}$$

then $E[Y|A, W]$ will not be estimated consistently. If a data-adaptive model selection algorithm is used to estimate $E[Y|A, W]$, the consistency of that estimate will rely on the ability of the algorithm to capture the true relationship. An algorithm based on linear combinations of tensor

DOI: 10.1002/sim

products, for example, would be unlikely to adequately capture a relationship that involves sharp discontinuities, as one would observe in the context of indicator functions, for example.

In contrast to the G-computation estimator, the targeted maximum-likelihood estimator is consistent if *either* $E(Y|A,W)$ *or* $P(A|W)$ is consistently estimated. Consistent estimation of $P(A|W)$ again relies either on the correct specification of an *a priori* model or the use of an algorithm that is well suited to the particular data-generating distribution at hand. The added robustness of the targeted maximum-likelihood approach is particularly valuable in contexts where the dependence of the biomarker on covariates is easier to model than the dependence of the outcome on biomarker and covariates.

While van der Laan and Rubin [1] provide both general theoretical proofs and practical illustrations of these improved robustness properties of the targeted maximum-likelihood approach, we briefly illustrate them here in the context of a simple simulation study that centers on estimating the VIM of a binary treatment variable A on a continuous outcome Y, adjusted for four putative confounders $W=(W_1,W_2,W_3,W_4)$. The data were generated as follows: W consists of four independent Bernoulli variables with $P(W_j=0)=P(W_j=1)=0.5$. The treatment variable A depends on W according to the logistic model

$$\text{logit}[P(A|W)]=\log\left(\frac{P(A|W)}{1-P(A|W)}\right)=-1+2W_1+W_2-W_3 \tag{10}$$

The outcome Y, finally, is normally distributed with standard deviation 0.2 and mean given by

$$E[Y|A,W]=-2+0.3A-0.5W_1-0.6W_2+0.4W_3+0.4W_4 \tag{11}$$

Based on (11), the true W-adjusted VIM of A on Y is given by

$$E_W(E(Y|A=1,W)-E(Y|A=0,W))=0.3 \tag{12}$$

We examined different approaches for estimating this VIM based on the limit that the approach would converge to as the sample size grows larger and larger. For this purpose, we applied each approach to a data set of 1 000 000 observations generated according to the rules above. The unadjusted VIM estimate is given by the estimate of the coefficient α_1 in the model

$$E[Y|A]=\alpha_0+\alpha_1 A \tag{13}$$

Fitting this linear regression model to the test data set of 1 000 000 observations, we obtained a limit of -0.101 for this approach. The unadjusted VIM is thus heavily confounded by W and does not even capture the right direction of the effect of A on Y. We next considered four different targeted maximum-likelihood estimators of the W-adjusted VIM that vary in the models they rely on for estimating $E[Y|A,W]$ and $P(A|W)$. Based on the simulation setup, the correct models are given by

$$E[Y|A,W]=\beta_0+\beta_1 A-\beta_2 W_1-\beta_3 W_2+\beta_4 W_3+\beta_5 W_4 \tag{14}$$

and

$$\text{logit}[P(A|W)]=\gamma_0+\gamma_1 W_1+\gamma_2 W_2+\gamma_3 W_3 \tag{15}$$

We also considered the mis-specified models

$$E[Y|A,W]=\beta_0+\beta_1 A \tag{16}$$

DOI: 10.1002/sim

Table I. First simulation study. Limit of the targeted maximum-likelihood estimator of the W-adjusted variable importance of A on Y for different models of the two nuisance parameters $E[Y|A,W]$ and $P(A|W)$.

Estimator	$E[Y\|A,W]$	logit$[P(A\|W)]$	Limit
1	$\beta_0+\beta_1 A+\beta_2 W_1+\beta_3 W_2+\beta_4 W_3+\beta_5 W_4$	$\gamma_0+\gamma_1 W_1+\gamma_2 W_2+\gamma_3 W_3$	0.300
2	$\beta_0+\beta_1 A+\beta_2 W_1+\beta_3 W_2+\beta_4 W_3+\beta_5 W_4$	$\gamma_0+\gamma_1 W_1$	0.300
3	$\beta_0+\beta_1 A$	$\gamma_0+\gamma_1 W_1+\gamma_2 W_2+\gamma_3 W_3$	0.300
4	$\beta_0+\beta_1 A$	$\gamma_0+\gamma_1 W_1$	0.068

The true variable importance is given by 0.3; the true model for $E[Y|A,W]$ is $\beta_0+\beta_1 A+\beta_2 W_1+\beta_3 W_2+\beta_4 W_3+\beta_5 W_4$; the true model for logit$[P(A|W)]$ is $\gamma_0+\gamma_1 W_1+\gamma_2 W_2+\gamma_3 W_3$. The limits shown are realizations of the estimators on a data set of 1 000 000 observations.

and

$$\text{logit}[P(A|W)]=\gamma_0+\gamma_1 W_1 \tag{17}$$

Table I shows that the targeted maximum-likelihood estimator converges to the true VIM of 0.3 as long as at least one of the two models for $P(A|W)$ and $E[Y|A,W]$ is correctly specified. Estimator 1 employs correctly specified models for both $P(A|W)$ and $E[Y|A,W]$ and thus, not surprisingly, converges to the truth. Estimators 2 and 3, however, converge to the true VIM of 0.3 in spite of relying on mis-specified models for $P(A|W)$ and $E[Y|A,W]$, respectively, in each case by virtue of specifying the other model correctly. Only estimator 4, which is based on mis-specified models for both $P(A|W)$ and $E[Y|A,W]$, fails to converge to the truth.

As mentioned above, in many applications the available subject-matter knowledge is not sufficient to allow an *a priori* decision about what models would be appropriate for $P(A|W)$ and $E[Y|A,W]$, leading the researcher instead to rely on various data-adaptive model selection algorithms. In simple cases as the one presented above, most such algorithms would have a good chance of identifying the appropriate functional form so that valid estimates of $P(A|W)$ and $E[Y|A,W]$ could be obtained. It is quite possible, however, that the true form of one of these two models is too complicated to be easily captured by these algorithms. In such cases, the ability to rely on a correct model for the other conditional distribution becomes extremely valuable.

We illustrate this point in a simple simulation study that follows exactly the same rules as above except that the conditional mean of Y given A and W is given by

$$E[Y|A,W]=-2+10AW_1W_2-8AW_3W_4 \tag{18}$$

The treatment variable A thus has a strong positive effect on Y if both $W_1=1$ and $W_2=1$, but also a strong negative effect if both $W_3=1$ and $W_4=1$. Since the W_j are independent of each other, each with mean 0.5, the true W-adjusted VIM of A is now given by

$$E_W(E(Y|A=1,W)-E(Y|A=0,W))=10\times 0.5\times 0.5-8\times 0.5\times 0.5=0.5 \tag{19}$$

The true dependence (18) of Y on A and W is now complicated enough to make it harder for many of the popular data-adaptive regression algorithms to capture it. Since the true model for $P(A|W)$ is still the same simple model as above, we might expect that conventional G-computation approaches will perform worse in this situation than targeted maximum-likelihood estimation, even if both approaches rely on data-adaptive model selection. We investigated this

DOI: 10.1002/sim

Table II. Second simulation study. Limit of four different data-adaptive G-computation estimators as well as the data-adaptive targeted maximum-likelihood estimator.

Estimator	Limit
G-computation (stepwise AIC)	−0.03
G-computation (D/S/A)	−0.03
G-computation (least angle regression)	−0.03
G-computation (polynomial splines)	−0.03
Targeted maximum likelihood	0.50

The true variable importance is given by 0.5. The limits shown are realizations of the estimators on a data set of 1 000 000 observations.

hypothesis by comparing the targeted maximum-likelihood estimator, based on the data-adaptive D/S/A algorithm as described above, with conventional G-computation estimators based on four different data-adaptive algorithms: stepwise selection based on Akaike information criterion [17], the D/S/A algorithm [13], least angle regression [12], and an algorithm based on polynomial spline functions [18]. Table II shows that the four G-computation estimators do in fact fail to converge to the true VIM of 0.50, with each of them converging to a value close to zero instead. The targeted maximum-likelihood estimator, on the other hand, still converges to the truth, thanks to its ability to rely on a less challenging model for $P(A|W)$.

5. RESULTS: IDENTIFICATION OF HIV MUTATIONS ASSOCIATED WITH DECREASED VIRAL SUSCEPTIBILITY TO LOPINAVIR

In this section, we present the results of applying three different approaches to assess the importance of each of a set of candidate PI mutations in determining clinical virologic response to lopinavir:

1. Estimation of the unadjusted association $E(Y|A=1)-E(Y|A=0)$ based on univariate regression of Y on A.
2. Estimation of the W-adjusted VIM $E_W(E(Y|A=1,W)-E(Y|A=0,W))$ based on the G-computation estimator (4).
3. Estimation of the W-adjusted VIM $E_W(E(Y|A=1,W)-E(Y|A=0,W))$ based on the targeted maximum-likelihood estimator (7).

Four hundred and one TCEs among 372 subjects involved initiation of a salvage regimen containing lopinavir and met all of our inclusion criteria. The frequency of the various candidate PI mutations among these TCEs is summarized in Table III. 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). As discussed in Section 3 and stated formally in equation (3) in Section 4, adjustment for covariates W requires that there be variation in the presence of the biomarker within the strata of W. In order to help ensure sufficient variation and the ability to control adequately for confounding, we estimated VIMs only for those mutations that occurred in at least 20 TCEs; among the mutations that had to be excluded based on this criterion are the important lopinavir resistance mutations 50V, 84C, and 88S. In addition, we assessed the extent of variation among the remaining mutations by examining the fitted probabilities $g_n^0(A,W)$. For a few of these mutations,

DOI: 10.1002/sim

Table III. Frequency of candidate protease inhibitor mutations among the 401 TCEs included in the analysis.

Mutation	Frequency	% Violations
10FIRVY	217	3
16E	9	—
20IMRTVL	115	0
23I	4	—
24IF	16	—
30N	45	64
32A	0	—
32I	21	58
33F	44	51
36ILVTA	141	0
46ILV	143	0
47V	17	—
48VM	16	—
48AST	1	—
50V	5	—
50L	0	—
53LY	33	0
54LMST	36	84
54VA	84	0
63P	311	5
71TVI	181	0
73CSTA	66	35
82AFST	100	6
82MLC	4	—
84AV	73	28
84C	2	—
88DTG	44	36
88S	9	—
90M	171	0

VIMs were estimated only for those mutations that occurred in at least 20 TCEs. For those mutations present in at least 20 TCEs, % Violations gives the percentage of TCEs with fitted mutation probabilities <0.05 or >0.95. These mutation probabilities reflect how likely a given mutation is to be present in a particular TCE, given the available baseline covariates capturing the patient's treatment history and the presence of mutations in the HIV reverse transcriptase gene. Fitted probabilities close to zero or one thus reflect that for a particular profile of baseline covariates the mutation of interest would almost always be absent or present, respectively. If such a lack of variation in the distribution of the mutation is observed in a high proportion of TCEs, the corresponding VIM estimates can become unreliable.

most notably 54LMST and 30N, a high proportion of the fitted probabilities were less than 0.05 or greater than 0.95, suggesting that they may not exhibit enough variation within the strata of W to allow for reliable VIM estimation. The results presented for these mutations should thus be interpreted with care.

It was not clear based on background knowledge whether the presence of mutations affected the duration until the outcome viral load was measured. We investigated this potential dependence

DOI: 10.1002/sim

Table IV. Estimated VIMs and associated p-values for candidate PI mutations.

		Unadjusted		G-comp		T-MLE	
Mutation	Score	VIM	p-value	VIM	p-value	VIM	p-value
10FIRVY	2	0.56	<0.01	0.28	0.12	0.26	0.30
20IMRTVL	2	0.46	0.02	0.39	0.04	0.37	0.04
30N	0	-1.09	<0.01	-0.60	0.03	-0.20	0.72
32I	10	0.80	0.01	0.63	0.03	0.81	<0.01
33F	5	0.83	<0.01	0.49	0.05	1.12	0.02
36ILVTA	1	0.29	0.10	0.39	0.03	0.39	0.04
46ILV	11	0.44	0.01	0.18	0.32	0.13	0.60
53LY	3	0.54	0.04	0.32	0.28	0.32	0.33
54LMST	10	0.67	0.01	0.15	0.55	0.16	0.72
54VA	11	0.86	<0.01	0.69	<0.01	0.61	<0.01
63P	2	0.10	0.57	-0.02	0.90	-0.07	0.72
71TVI	2	0.34	0.03	0.24	0.13	0.24	0.17
73CSTA	2	0.79	<0.01	0.61	0.02	0.46	0.36
82AFST	20	0.68	<0.01	0.49	0.02	0.64	<0.01
84AV	11	0.50	0.02	0.25	0.19	0.49	0.04
88DTG	0	-0.86	<0.01	-0.50	0.05	-0.37	0.33
90M	10	0.52	<0.01	0.45	0.02	0.45	0.02

Score refers to the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 18 July 2006). The p-values shown are based on 10 000 bootstrap samples.

by using box plots to compare the distribution of outcome monitoring times in the presence versus absence of each mutation. These plots did not suggest any major differences in the distribution of monitoring times according to the presence or absence of any mutation. In addition, we fit a data-adaptive model of the conditional hazard of viral load monitoring over time in order to examine the potential dependence of monitoring on the presence of candidate mutations and baseline covariates. The data-adaptively selected model included as single covariate the time that had elapsed since initiation of the new treatment regimen. Together, these findings suggest that the presence of particular mutations did not strongly affect monitoring time, reducing concern regarding the assumption that mutation effect was constant over time (discussed in Section 3.2).

Table IV summarizes the unadjusted associations and estimates of the W-adjusted VIM based on the G-computation and targeted approaches, along with associated p-value. Table V gives three different rankings for the set of candidate mutations, based on the p-values generated by each of the three approaches. The mutation ranking generated by the current Stanford scoring system is included for comparison. Inference was based on non-parametric bootstrap sampling, respecting the subject rather than the TCE as the independent unit of analysis. The resulting p-values were adjusted for multiple testing using the Benjamini–Hochberg method [19] to control the false discovery rate (aiming to ensure that the expected proportion of false positives was 0.05).

Among the 17 candidate PI mutations considered here, the Stanford scoring system identifies the following seven mutations as major contributors to lopinavir resistance: 82AFST, 54VA, 46ILV, 84AV, 90M, 32I, and 54LMST; the remaining 10 mutations are thought to make minor or no contributions to resistance. The unadjusted association analysis yielded significant p-values for all but two of the candidate PI resistance mutations (36ILV and 63P). The significant subset thus included eight mutations thought to have a minor or no effect on lopinavir resistance. Among these

DOI: 10.1002/sim

Table V. Candidate PI mutations ranked according to the p-values of three distinct VIM estimates.

Score		Unadjusted		G-comp		T-MLE	
Mutation	Score	Mutation	p-value	Mutation	p-value	Mutation	p-value
82AFST	20	30N*	<0.001	54VA	<0.001	82AFST	0.001
54VA	11	54VA	<0.001	82AFST	0.018	54VA	0.003
46ILV	11	82AFST	<0.001	90M	0.019	32I	0.003
84AV	11	33F	<0.001	73CSTA	0.019	90M	0.024
90M	10	10FIRVY	0.001	32I	0.033	33F	0.024
32I	10	73CSTA	0.001	30N*	0.033	36ILVTA	0.035
54LMST	10	88DTG*	0.001	36ILVTA	0.034	84AV	0.037
33F	5	90M	0.003	20IMRTVL	0.043	20IMRTVL	0.039
53LY	3	32I	0.014	33F	0.051	71TVI	0.174
10FIRVY	2	46ILV	0.015	88DTG*	0.051	10FIRVY	0.301
73CSTA	2	54LMST	0.015	10FIRVY	0.123	53LY	0.330
20IMRTVL	2	84AV	0.016	71TVI	0.130	88DTG*	0.330
71TVI	2	20IMRTVL	0.016	84AV	0.193	73CSTA	0.361
63P	2	71TVI	0.034	53LY	0.277	46ILV	0.600
36ILVTA	1	53LY	0.039	46ILV	0.321	63P*	0.719
30N	0	36ILVTA	0.097	54LMST	0.551	30N*	0.719
88DTG	0	63P	0.574	63P*	0.898	54LMST	0.719

Score refers to the resistance score assigned to a mutation by the Stanford HIVdb scoring system (accessed on 18 July 2006). The p-values shown are based on 10 000 bootstrap samples. Mutations marked with an asterisk have a negative VIM estimate, suggesting that they contribute to an improved rather than diminished virologic response.

were the mutations 30N and 88DTG, both estimated to be significantly protective. The protective association of 30N with the outcome was in fact ranked the most important of the unadjusted associations. In addition, multiple mutations considered by the current knowledge to have only minor effects on resistance (for example, 33F, 10FIRV, and 73CST) ranked higher than most of the known major lopinavir resistance mutations (such as 90M, 32I, and 54LMST).

After adjusting for covariates using G-computation, fewer mutations were identified as significant, and the resulting ranking agreed to a greater extent with the current knowledge. Specifically, this approach identified eight mutations as having a significant impact on lopinavir resistance, with an additional two mutations found to be borderline significant (p-values of 0.051 for 33F and 88DTG). This group of 10 mutations includes both four of the seven major lopinavir resistance mutations and six mutations thought to make minor or no contributions to resistance. In particular, we note that the mutations 30N and 88DTG were still identified as having a protective effect.

Targeted maximum-likelihood estimation of the adjusted VIM provided the ranking in best agreement with the current knowledge. The significant subset of mutations identified by this approach included five of the seven major known mutations and only three minor mutations (33F, 36ILV, and 20IMRTV). The mutation considered most important for lopinavir resistance, 82AFST, was ranked highest, followed by three major known lopinavir resistance mutations (32I, 54AV, and 90M). Unlike G-computation, targeted maximum likelihood also identifies the major lopinavir resistance mutation 84AV as a significant contributor to resistance. In addition, unlike the other two approaches, it did not rank either 88DTG or 30N as significantly protective. Two mutations thought to be important for lopinavir resistance, 46ILV and 54LMST, were not identified by targeted VIM

DOI: 10.1002/sim

estimation. However, Table III reports that for the mutation 54LMST, 84% of observations had fitted mutation probabilities <0.05 or >0.95, suggesting a lack of variation in 54LMST within the strata of W, which may lead to unreliable VIM estimates. In addition, *in vitro* experiments examining the effect of 46ILV on viral phenotype suggest that this mutation may in fact be less important for lopinavir resistance than previously thought [20].

6. DISCUSSION

6.1. HIV resistance mutations

The current article discussed how targeted maximum-likelihood estimation of VIMs can be used in biomarker discovery. Motivation for the method, details of its implementation, and interpretation of results were illustrated using an example from the treatment of HIV infection. We estimated the importance of each of a set of candidate PI mutations for clinical virologic response to treatment with the commonly used PI drug lopinavir, adjusted for covariates including treatment history, the presence of non-PI mutations, and characteristics of the background regimen.

Our analysis suggests that targeted maximum-likelihood estimation of VIM represents a promising new approach for studying the effects of HIV mutations on clinical virologic response to antiretroviral therapy. The subset of mutations identified by this approach as significant contributors to lopinavir resistance was in better agreement with the current knowledge than the subsets identified by an unadjusted analyses or the G-computation approach. Specifically, the unadjusted analysis identified as significant all but two of the candidate mutations, including eight mutations thought to have a minor or no effect on lopinavir resistance. G-computation reduced the significant subset to four of the seven mutations thought to make major contributions to lopinavir resistance, while still including six mutations thought to make only a minor or no contribution to resistance. In contrast, the significant subset of mutations identified by targeted maximum likelihood included five of the seven major known mutations and only three minor mutations. In addition, the specific ranking provided by targeted VIM estimation also agreed better with the current understanding than did the rankings generated with alternative methods.

While targeted VIM estimates were able to replicate most known findings, they also suggested that the mutation 46ILV may be less important in determining resistance to lopinavir than previously thought. As mentioned in Section 5, this finding has some support from *in vitro* studies [20], suggesting that a more detailed investigation of the role of this mutation may be warranted. Taken as a whole, the promising results reported here suggest that further application of the targeted VIM approach may result in improvements to existing genotypic interpretation algorithms.

6.2. Targeted maximum likelihood

As illustrated in this article, targeted maximum-likelihood estimation offers an improvement in robustness over conventional likelihood-based approaches, which is straightforward to implement using standard statistical software. Specifically, the approach remains consistent if we mis-specify how virologic response depends on the mutation and all covariates, but correctly model how the presence of the mutation depends on covariates. The resulting targeted VIM estimates provide a means to both rank candidate biomarkers and to identify a subset of biomarkers as relevant for a given outcome. The current article focused primarily on VIM for a continuous outcome. Generalization to a binary outcome modeled using logistic regression is straightforward, as was

DOI: 10.1002/sim

mentioned briefly. The method can further be generalized to alternative approaches for obtaining an initial estimate of $E(Y|A, W)$.

The double robust variable importance estimator introduced by van der Laan [2] provides similar advantages to the targeted VIM estimate in terms of improved robustness to model misspecification. However, the targeted approach has several practical advantages. Many practitioners are more familiar with regression-based approaches, as used by the targeted estimator, than with the estimating function methodology employed by the double robust estimator. In addition, the targeted maximum-likelihood VIM estimator can in many cases be implemented using standard software, in a natural extension of common regression approaches. These practical advantages, together with the improvement in robustness, make targeted maximum-likelihood estimation of variable importance a promising new approach to biomarker discovery.

REFERENCES

1. van der Laan M, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2006; **2**(1):Article 11. Available from: http://www.bepress.com/ijb/vol2/iss1/11.
2. van der Laan M. Statistical inference for variable importance. *The International Journal of Biostatistics* 2006; **2**(1):Article 2. Available from: http://www.bepress.com/ijb/vol2/iss1/2.
3. Robins J. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy survivor effect. *Mathematical Modelling* 1986; **7**:1393–1512.
4. Brun-Vezinet F, Descamps D, Ruffault A, Masquelier B, Calvez V, Peytavin G, Telles F, Morand-Joubert L, Meynard JL, Vray M, Costagliola D, NAS Group. Clinically relevant interpretation of genotype for resistance to abacavir. *AIDS* 2003; **17**(12):1795–1802.
5. Laethem KV, Vaerenbergh KV, Schmit J, Sprecher S, Hermans P, Harrer VDVRST, Witvrouw M, Wijngaerden EV, Stuyver L, Ranst MV, Desmyter J, Clercq ED, Vandamme A. Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed hiv-1 genotypic populations. *Journal of the Acquired Immunodeficiency Syndrome* 1999; **22**(2):107–118.
6. Shafer R, Jung D, Betts B. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nature Medicine* 2000; **6**(11):1290–1292.
7. Hammer S, Saag M, Schechter M, Montaner J, Schooley R, Jacobsen D, Thompson MA, Carpenter C, Fischl M, Gazzard B, Gatell J, Hirsch M, Katzenstein D, Richman D, Vella S, Yeni P, Volberding P, IA Society USA Panel. Recommendations of the International AIDS Society—USA Panel. *Journal of the American Medical Association* 2006; **296**(7):827–843.
8. Robins J, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; **3**:143–155.
9. van der Laan M, Robins J. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer: Berlin, 2003.
10. Breiman L, Friedman JH, Olshen R, Stone CJ. *Classification and Regression Trees*. The Wadsworth Statistics/Probability Series. Wadsworth International Group: Belmont, CA, 1984.
11. Breiman L. Random forests. *Machine Learning* 2001; **45**:5–32.
12. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *The Annals of Statistics* 2004; **32**(2):407–499.
13. Sinisi S, van der Laan M. Deletion/Substitution/Addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1):Article 18. Available from: www.bepress.com/sagmb/vol3/iss1/art18.
14. Bickel P, Klaasen C, Ritov Y, Wellner J. *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press: Baltimore, MD, 1993.
15. Bembom O, van der Laan M. A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electronic Journal of Statistics* 2007; **1**:574–596.
16. Bembom O, Fessel W, Shafer R, van der Laan M. Data-adaptive selection of the adjustment set in variable importance estimation. *Technical Report 231*, *UC Berkeley Division of Biostatistics Working Paper Series*, 2007. Available from: http://www.bepress.com/ucbbiostat/paper231.
17. Venables W, Ripley B. *Modern Applied Statistics with S*. Springer: Berlin, 2002.

DOI: 10.1002/sim

18. Kooperberg C, Bose S, Stone C. Polychotomous regression. *Journal of the American Statistical Association* 1997; **92**:117–127.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B* 1995; **57**:289–300.
20. Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences* 2006; **103**: 17355–17360.

DOI: 10.1002/sim